Scope
Volume 14 Number 03 September 2024

Performance Analysis of World Happiness using Machine Learning Algorithms

Dr. S. Lakshmi
Computer Applications, SRM Institute of Science and Technology, Chennai, Tamilnadu, India
Orchid id: 0000-0003-2553-778X

Abstract: Happiness is the ultimate goal of each and every individual. It is very challenging task to define the happiness and measure the happiness since the perception of happiness may differ for everyone. Phycologists are trying to assess the happiness by considering some key signs such as “Feeling satisfied”, “Enjoying the positive and healthy relationship with others”, “Experiencing gratitude”, “Practicing kindness”, “Being with loved ones or friends” and “Willingness to take challenges”. The purpose of this paper is to conduct the study on world happiness report dataset and find out the main features of happiness of the human beings. It has been identified that the GDP per capita is the main factor for happiness. The second choice is the life span of the human being. The GDP and Life expectancy played a major role to decide the happiness score and this could be the general opinion of all human without any doubt and hesitation. Being healthy and wealthy people are all happy in their real life? In this paper, I try to find out answer for this question by considering other factors of the world happiness index dataset.

Keywords: Machine learning algorithms, Happiness index, Performance Evaluation

I. Introduction

According to Aristotle there are two different types of happiness. They are Hedonia and Eudaimonia. Hedonia is originated from pleasure which means “doing things which makes you feel good”, “fulfilling the desires of your loved ones”, “feeling of enjoyment”, “experiencing something which gives enjoyment”, “feeling of satisfaction” and “contentment”. Ryan & Deci [1] defined the hedonism as the pleasure over pain which dies the reward and punishment. It has three main components such as life satisfaction, the presence of a positive mood and the absence of a negative mood. Eudaimonia is a type of happiness which derived from finding the meaning of life, purpose of life and value of life. It depends on fulfilling the responsibilities. Eudaimonia is the mixture of eu which means good or well and daimon which means spirit. Eudaimonia also means as welfare or flourishing. According to Socrates and Plato virtue is the form of knowledge like courage, justice and so on.

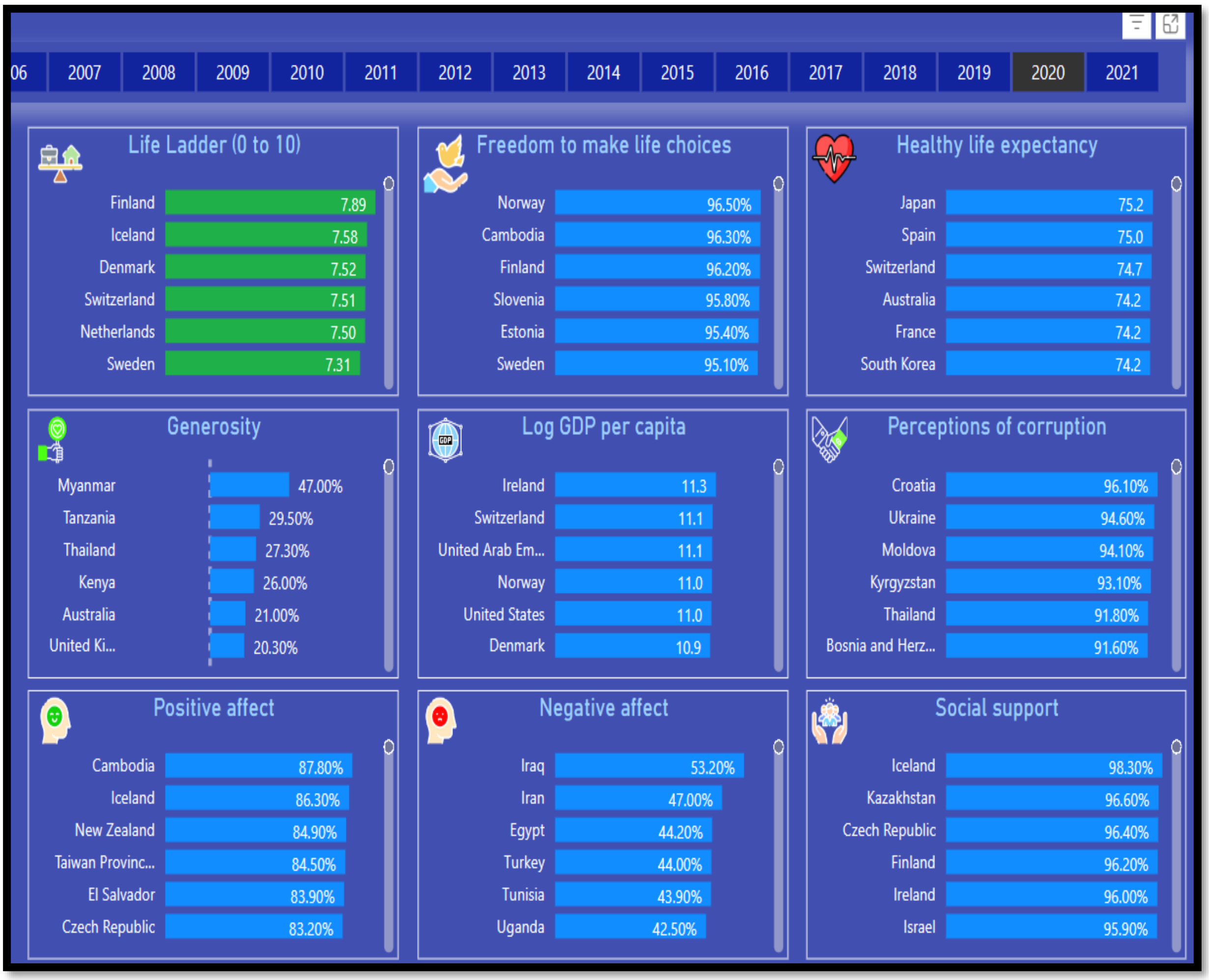
Most of the individuals define the happiness and well-being as the main purpose of our life. Generally, the happiest mindset leads to lower the blood pressure and reduce the stress. It also improves the diet, sleeping hours and make us to do exercise properly and regularly. It provides the ability to solve the problems efficiently and make us to think optimistic. In this world happiness report [2] consider the features such as Health life, GDP per capita, Freedom, Family, Destopia, Trust, Corruption and generosity to calculate the happiness index by taking more than 150 countries to measure the happiness. The relationship between happiness and wellbeing and the differences between happiness and other things such as competence, positive emotions discussed in [3]. The happiness of the people will differ from time to time and one season to another. We could not forget the situation during COVID-19 period. Almost we lost our freedom and lead our life with full of fear [4]. The status of World happiness described in detail during COVID-19 pandemic [5]. The physical and mental needs including education the nation's happiness index calculated [12]. In [13], the Myers–Briggs personality theory used to analyse the indicators like HDI, GDP per capita, and democracy index. In [14], the World Happiness Index using regression analysis and correlation to study calculation issues related to this index based on seven components of this global indicator for 156 countries collected from the 2016 World Health Report. The machine learning algorithms used to develop prediction models such as NB, K-nearest neighbor (KNN), MLP, and Decision Tree (DT), based on survey data collected from employees of the Ministry of Public Health in Thailand [15].

By considering all the factors in the world happiness dataset, identify the best machine learning algorithm with highest accuracy in happiness prediction is the aim of this paper.

II. Materials and Methods

The world happiness survey is being conducted every year and the details are updated. The Fig:1 shows the survey index report for the year 2020 which is available in the Kaggle website[11]. This report is prepared by using Microsoft powerBI data visualization technique. According to this survey, Finland stands first with 7.89 happiness index score. Norway has 96.5% in freedom, Myanmar has 47% in generosity and Iceland has 98.3% in social support. All these indicators played a major role to decide the happiness index. Commonly the GDP has got the highest priority to decide the happiness index. Life expectancy has got the second place to decide the happiness index.

Fig 1: Analysis of the Happiness index Report 2020



In this work the relationship between various happiness indicators and their impact on happiness discussed in detail. The entire data analysis is done using statistical and machine learning algorithms to calculate the happiness index and the process is depicted diagrammatically in the flow diagram of Fig.2.The workflow of this happiness index calculation process starts by combining the world happiness dataset, human development index dataset and the crime index dataset and the pre-processing step follows the merging process. The correlation coefficient is calculated under the statistical process then the various machine learning algorithms are applied and the model building is done. The performance of the model is evaluated and the results are tabulated.

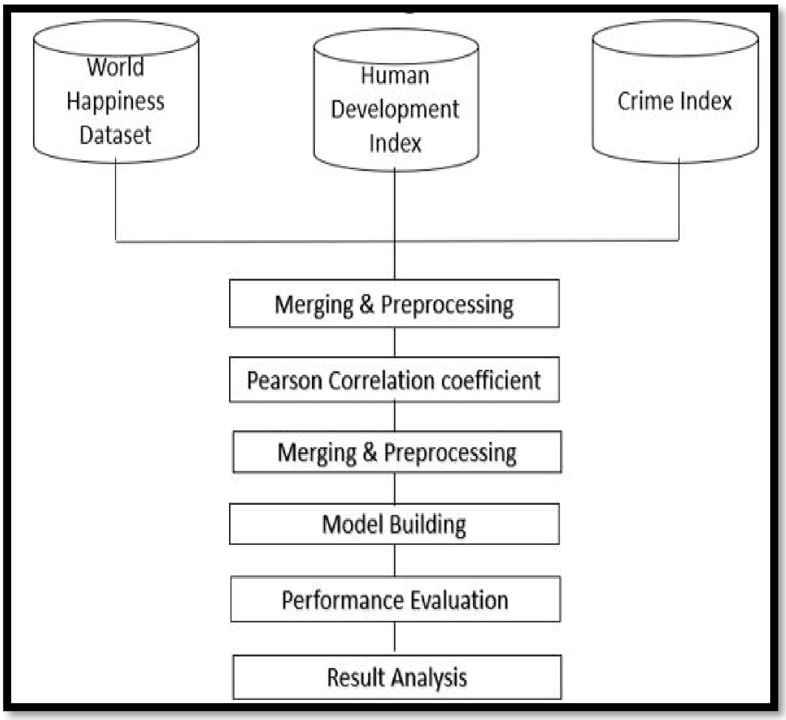


Fig:2 Proposed System

2.1 Dataset

This work is done based on three datasets named Human development index, World Happiness index, and the Crime index dataset. The Kaggle world happiness dataset which has the following features Country, Region, Happiness Rank, Happiness Score, Standard Error, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual to identify the happiness index. Nearly the dataset has the details about 150 countries. The Happiness score is collected from respondents and it ranges from 0 to 10. The features are added in this dataset by getting responses from various level of people through survey by considering the day-to-day experiences. The data sample is shown in the table 1 by taking the first 10 rows from the dataset using head () function in python. The datasets with few fields are listed for sample in Tables 1,2 and 3.

Table: 1.1 World Happiness index dataset

Country	**Happiness Rank**	**Happiness Score**	**Economy (GDP per Capita)**	**Family**	**Health (Life Expectancy)**	**Freedom**	**Dystopia Residual**
Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	2.51738
Iceland	2	7.561	1.30232	1.4022	0.94784	0.62877	2.70201

				3			
Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	2.49204
Norway	4	7.522	1.459	1.33095	0.88521	0.66973	2.46531
Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	2.45176
Finland	6	7.406	1.29025	1.31826	0.88911	0.64169	2.61955
Netherlands	7	7.378	1.32944	1.28017	0.89284	0.61576	2.4657
Sweden	8	7.364	1.33171	1.28907	0.91087	0.6598	2.37119
New Zealand	9	7.286	1.25018	1.31967	0.90837	0.63938	2.26425
Australia	10	7.284	1.33358	1.30923	0.93156	0.65124	2.26646

Table 1.2: Human Development Index dataset

		Human Development Index (HDI)	Life expectancy at birth	Mean years of schooling	Gross national income (GNI) per capita	Adjustment factor for planetary pressures	Carbon dioxide emissions per capita (production)	Carbon dioxide emissions (production) index
1	Switzerland	0.962	84	13.9	66,933	0.828	3.7	0.946
2	Norway	0.961	83.2	13	64,660	0.764	7.6	0.889
3	Iceland	0.959	82.7	13.8	55,782	0.66	8.6	0.875
4	Hong Kong, China (SAR)	0.952	85.5	12.2	62,607	..	4.2	0.939
5	Australia	0.951	84.5	12.7	49,238	0.67	15.4	0.776
6	Denmark	0.948	81.4	13	60,365	0.847	4.5	0.934
7	Sweden	0.947	83	12.6	54,489	0.848	3.8	0.944
8	Ireland	0.945	82	11.6	76,169	0.722	6.8	0.902

9	Germany	0.942	80.6	14.1	54,534	0.854	7.7	0.888
10	Netherlands	0.941	81.7	12.6	55,979	0.791	8.1	0.883
11	Finland	0.94	82	12.9	49,452	0.777	7.1	0.897
12	Singapore	0.939	82.8	11.9	90,919	0.709	7.8	0.887
13	Belgium	0.937	81.9	12.4	52,293	0.792	7.2	0.895
13	New Zealand	0.937	82.5	12.9	44,057	0.807	6.9	0.899

Table 1.3: Safety index dataset

Abu Dhabi, United Arab Emirates	11.67	88.33
Doha, Qatar	13.96	86.04
San Sebastian, Spain	14.86	85.14
Taipei, Taiwan	15.05	84.95
Quebec City, Canada	15.14	84.86
Ajman, United Arab Emirates	15.64	84.36
Sharjah, United Arab Emirates	15.69	84.31
Dubai, United Arab Emirates	16.3	83.7
Zurich, Switzerland	17.26	82.74

The happiness index score is divided into four categories based on the index value by using the variability measurement in statistics. Two methods are used in this paper. In the first method, the happiness index values are added in the order and the lowest value is subtracted from the highest value to calculate the range. This is considered as the most informative measure of variable data. In this data set, the highest value of the happiness index is 7.5087 and the lowest value is 2.5669. The difference is 5.2418. Since we have four intervals, the range is divided by 4 and we get 1.131. In the second method, the standard deviation is calculated and the value of std is 1.123. By using these values, the range is fixed for all four intervals as in the Table.4.

Table 4: Happiness index score intervals

S.No	**Interval**	**Happiness score range**
1	Unhappy	2.57 to 3.88
2	Moderately happy	3.88 to 5.19
3	Happy	5.19 to 6.5
4	Very happy	6.5 to 7.81

Hence the happiness index is divided into four categories and the range for each and every category is tabulated. The human development report produced by the Human Development Report office of the United Nations Development Programme (UNDP). The Human development index can be measured by considering the HDI score, GNI per capita, gender development, gender inequality, life expectancy, number of years of schooling, mean years of schooling, inequality in education, planetary pressure, carbon-dioxide emission, mortality ratio, adolescent birth rate and labor data. The third data set consists of the Safety index of all the countries as well as the crime index. The crime index is based on the kidnapping rate, robbery rate, imprisonment rate and theft rate. These three datasets provide us the concrete base for understanding the happiness index.

2. 2. Proposed Methodology

The proposed methodology consists of the following steps: Data acquisition, Data Preprocessing, Data visualization, splitting the dataset, apply various machine learning algorithms for training the model, then the comparison of results through various metrics.

2.2.1. Data collection

The dataset acquisition process started by merging the Happiness development index, the world happiness index and the crime and safety dataset for further processing.

2.2.2. Data Pre-processing

Data clearing is the prime step in pre-processing in which all missed data and unwanted noisy data could be eliminated. The correlation between the attributes so that we are able to find out the relevant features for identifying and extracting the features for further processing. The standard scalar technique is the normalization technique used to standardize the data set which is in different scales for improving the model performance.

2.2.3. Model building and Analysis

Various machine learning algorithms are used to build the model such as Linear regression, Lasso regression, XGBoost algorithm, Lasso regression, SVR and Decision tree. The models are evaluated by using the metrics such as R-squared, mean square error, mean absolute error and root mean squared error. The analysis used to give the reliable outcome which can be used to frame the policy goals and the determinants of happiness life index.

1. Linear Regression

   It is considered as a basic model and the equation for this machine learning model is

   $y = mx + c$ -----------------------------(1)

   where y is the dependent variable and x is an independent variable m is the estimated parameter of slope and c is the intercept used to represent errors in the form of matrix.

2. Multiple linear Regression

According to the report by [9], machine learning algorithms produce imperfect results due to considering few features while clustering.

This model can use multiple independent variables to predict a single dependent variable[6]. The general equation of the multiple linear regression is

$$y = b0 + b1x1 + b2x2 + \cdots .. + bpxp \text{ ---------------------------(2)}$$

3.Support Vector Machine

It is a regression algorithm. It follows the SVM's principle. Using the hyperplane as a separating option in different classes.

4.Decision Tree

This is a tree structure machine learning method. The internal node is a parent node or a judgement node and the leaf node consist of the judgement result which is the predicted values of the target variables.

5.Random Forest

Ensemble methods are machine learning methods used to produce the predictive models to get more accurate results. Random forest is an ensemble algorithm which is based on the decision trees. Several trees are combined together to form a random forest using bagging technique.

6. XGBoost regression: It is the Extreme Gradient Boost which is tree based boosting machine learning algorithm.

7.AdaBoostRegressor is also an ensemble technique which trains the model sequentially.

2.2.4. Model evaluation

To analyse and evaluate the model, the error and goodness and fit are calculated. Here mean square error, root mean square error and the mean absolute errors can be calculated and the performance of the model is evaluated.

Root Mean Square Error (RMSE): It is used to show the standard deviation between the predicted and absolute value. The formula for calculating the root mean square error is

$$RMSE = \sqrt{}(\sum_{k=0}^{n} \frac{(\mathrm{yi-yp})2}{\mathrm{n}})\text{----------------------------(3)}$$

Mean Square Error (MSE): Generally, the errors are the differences between the predicted values and the actual values of the variables. They can be calculated by using the formula

$$MSE = |Yi - Yp|/n \text{ --------------------------------(4)}$$

where

yi = actual values
yp = predicted values
n = number of observations /rows

Mean Absolute Error (MAE)

This is calculated by taking the absolute difference between the actual values and the predicted values. The formula for calculating the MAE is as follows:

$MAE = 1/n\sqrt{}|y-(pred(y)|$----------------------------(5)

III. Results and Discussion

Quality of life played a major role in the happiness index score prediction. To predict the quality-of-life various indicators are combined from various domains like social, economic and health[10]. Measuring or assessing the quality of life and happiness index is not at all an easy task and also it is a challenging task too. The supervised machine learning approaches used in the dataset for prediction. Naturally the two data fields ie., GDP and Health life expectancy played a vital role to lead a happy life as per the most of the views of the people in the survey. For classifying the happiness score the common classification algorithms such as Naïve Bayes classifier, Nearest neighbour, Support vector machine classifier and Decision tree can be used. Since the Gross per capita income alone is not sufficient to calculate the happiness index of the human being.

Table 5: Pearson Correlation Coefficient for World happiness index

Statistical technique **Pearson Correlation**	**Correlation coefficient**
Economy GDP Per Capita	0.780965527
Family	0.740605197
Health (Life Expectancy)	0.724199595
Freedom	0.568210904
Trust (Government Corruption)	0.395198584
Generosity	0.180318527
Dystopia Residual	0.530473518

Table 5 shows the Pearson correlation coefficient for all the indicators with the happiness score indicator. The table 6 shows the correlation coefficient of the dependent values by using the dependent variable as happiness index with all the independent variables in the dataset. In Table 7 shows the correlation among the multiple independent variables with the dependent variable ie., happiness index and the values are tabulated clearly. The various regression analysis methods are used and the performance metrics are tabulated in Table 8.

Table 6: Linear Regression with correlation coefficient

Algorithm- Linear Regression Dependent variable is Happiness index	Correlation Value
Economy GDP	86.12
Family	86.56
Health	70.12
Freedom	73.33
Trust	72.21
Generosity	35.27
Dystopia residual	30.24

Table 7: Multiple Linear Regression with correlation coefficient

Algorithm – Multiple linear Regression Dependent variable : Happiness index	Correlation coefficient
Economy& Family	86.11
Economy, Family& Health	85.94
Economy, Family& Health, Dystopia Residual	85.96
Economy, Family& Health, Dystopia Residual, Freedom, Generosity and Trust	84.68

Table 8: Comparison of various proposed regression Models

Algorithm	Performance Evaluation			
	MAE	MSE	RMSE	R2 Score
Linear Regression	0.053	0.0045	0.0657	0.75
Multiple Linear Regression	0.054	0.004	0.067	0.765
Decision Tree Regression	0.047	0.0052	0.071	0.91(3)
Random Forest Regression	0.064	0.0045	0.08	0.925(1)
Support vector Regression	0.067	0.006	0.08	0.887
XGBoost Regression	0.066	0.0048	0.09	0.921(2)
Gradient Boosting	0.065	0.0046	0.127	0.83
AdaBoosting	0.09	0.003	0.108	0.843

The performance metrics are calculated for all the models and the values are tabulated. From the analysis of the Random Forest Regression algorithm produces the R2 score as 0.925 which is the highest level of performance. The R2 score of the XGBoost regression technique is 0.921 and its Mean square error is 0.0048. The third place is taken by the decision tree regression which has the R2 value of 0.91 and the MSE is 0.0048.

In the Linear regression the MAE is 0.05, MSE is 0.0045 and RMSE is 0.0657 and the r2 score is 0.75 which is very less when we compare with random forest, support vector and decision tree algorithms. Even through the linear regressor produces better results for world happiness index calculation, its performance is comparatively less when we combine the three datasets to produce the happiness index calculation.

Hence the happiness index is calculated based on the GDP income which shows the highest weightage in world happiness index dataset as well as the human development index. According to World happiness index of the top 10 countries having highest happiness index is Luxembourg, Singapore, Ireland, United Arab Emirates, Kuwait, Norway, Switzerland, Hong Kong, S.A.R. of China and Unites states of America. Some of the questions are listed above while calculating the happiness index.

1. Can we say that the people happily leads their life in these ten countries?
2. Can GDP alone give satisfaction to the human being?
3. Is unhealthy person can lead their life happily in these countries?
4. How can be the Unhealthy person happily leads their life?
5. If there is no freedom in the country, people can lead their life happily?
6. How the people can be happy in the corrupted environment?
7. What is the role of family in happiness index?
8. What is the role of education in happiness?
9. How generosity is related to happiness?
10. How the crime index affects the happiness index?

Finland stands first in happiness index i.e., 7.8 and healthy life expectancy in is 82.48 years in Finland. The GDP of Finland is high, life expectancy is also high.

In Singapore the healthy life expectancy is 77% and the life expectancy is also good. Hence people always prefer to live in Singapore.

Myanmar stands first in Generosity and the happiness index is 4.39 and the happiness rank is 115. Since the index is 4.39, it comes in moderate. Even through the Myanmar is the top scorer in humanity and the happiness rank is at last. Does it mean the people in Myanmar living their life without happiness? When we correlate the political circumstances ie., corruption with happiness index, we can find solution for unhappiness. The environment, political situation, education, child health, literacy, safety index, freedom may be the reason for happiness.

According to Ovaska & Takashima [6], FP does not consider the hidden costs such as inflation and unemployment. The GDP consider only the natural capital, knowledge, health and social. Measuring happiness, therefore, should not only consider observable objective well-being measures (e.g., health and socioeconomic status), but also subjective well-being measures, such as domain satisfaction and quality of life. Easterlin [8] stated that the long-term monetary gains have small effects on quality of life.

When we combine the Happiness development index dataset with the world happiness index dataset, we are able to identify the other factors like education, no of years of schooling, carbon dioxide emission like the environmental factors to calculate the happiness index.

When we combine the crime index dataset with the happiness dataset, we are able to identify the safety environment so that we can calculate the happiness index by including the safest environment as an indicator.

When we combine all three datasets for calculating the happiness index, we are able to consider as many as indicators to decide the happiness index which will give the right direction to lead our life happily, safely and securely. We are able to identify where we are logging and how to improve to reach the state of happiness.

IV. Conclusion and Future Enhancement

The happiness index measurement is really a challenging task even though we have various datasets such as world happiness index dataset, human development index dataset and safety index dataset. Moreover, taking decisions by considering the GDP, Health life expectancy, crime index and literacy data. Most of the data collected through the survey by considering the levels such as dissatisfied, satisfied, neither satisfied nor dissatisfied and very much satisfied. To measure the satisfaction with work, we can consider the pay, productivity, interest, remuneration and growth and working hours as the indicators.

To measure the satisfaction with work, the following components played a key role. They are salary, work-life balance, productivity and interest in job. The GDP cannot differentiate the negative and positive impact of the wellbeing of the society. Generally the hidden costs of economic developments, inflation and the unemployment should not be included and discussed in the healthiness and wealth indicators of the happiness index. One more thing is mood swing which can play a major role to measure the mental status of the employees. In this work, to some extent, important features are considered for calculating the happiness index.

For future research, we need to collect more data related to subjective measures of the happiness and try to implement not only machine learning algorithms but also deep learning algorithms to calculate the happiness index for improving the quality of life. By using time-series algorithms for building models to calculate the happiness index is the future enhancement of this work. Further research plan is extending the work for collecting data from social media networks and analysing it for measuring the happiness index.

References:

1. Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52,* 141–166.

2. Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network
3. Theo Theobald, Cary Cooper, "The relationship between happiness and wellbeing", Doing the Right Thing, 2012, ISBN : 978-0-230-29844-6
4. GALLUP BLOG, MARCH 19, 2021, Did Happiness Survive the COVID-19 Pandemic?
5. Bansal P. The Ravaged Psyche: Impact of the COVID-19 Pandemic on the Human Mind. Hu Arenas. 2022;5(4):694–706.
6. Ovaska, T., & Takashima, R. (2006), Economic policy and the level of self-perceived well-being: An international comparison. The Journal of Socio-Economics, 35, 308–325
7. G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," Procedia - Soc. Behav. Sci., vol. 106, pp. 234–240, Dec. 2013
8. Stevenson, Betsey, and Justin Wolfers. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." Brookings Papers on Economic Activity, vol. 2008, (Spring,2008), pp. 1–87. JSTOR,
9. O. Spiga et al., "Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease," Orphanet J. Rare Dis., vol. 15, no. 1, p. 46, Dec.2020,
10. "World Happiness Report 2015-2021," Kaggle, March 2021. [Online]: [Accessed 5 10 2021].
11. Saputri T. R. D., Lee S. D,"A Study of Cross-National Differences in Happiness Factors Using Machine Learning Approach", International Journal of Software Engineering and Knowledge Engineering. 25(09n10), 1699–1702 (2015).
12. Yaman E., Music-Kilic A., Zerdo Z. "Using Classification to Determine Whether Personality Profiles of Countries Affect Various National Indexes", 2018 International Conference on Control, Artificial Intelli-gence, Robotics & Optimization (ICCAIRO). 48–52 (2018).
13. Carlsen L,"Happiness as a sustainability factor. The world happiness index: a posetic-based data analysis", Sustainability Science. 13 (2), 549–571 (2018).
14. Chaipornkaew P., Prexawanprasut T, "A Prediction Model for Human Happiness Using Machine Learning Techniques", 2019 5th International Conference on Science in Information Technology (ICSITech). 33–37(2019).