STUDY OF WORLD HAPPINESS REPORT DATASET USING MACHINE LEARNING METHODOLOGIES

***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 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 goal is to identify the elements that lead to happiness in various nations and make links to 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.

II. LITERATURE SURVEY

[1] **Title**: "World Happiness Report (2023)"
Authors: Andrew N. Rowan, Wellbeing International
Summary: Overall, the country's health has held up exceptionally well in spite of the rise in worldwide illness and mortality brought on by the pandemic and Russia's invasion of Ukraine. During the epidemic, there was a surge in helpful and altruistic actions, which may have mitigated the negative consequences of the pandemic and the lockdowns that followed.

[2] **Title**: “A Data Analysis of The World Happiness Index and Its Relation to The North-South Divide”
Authors: Charles Alba
Summary: One of the main goals is to show how qualitative data with bottom-up viewpoints on wellbeing are an essential supplement to quantitative self-report measures, enabling more complex cultural understandings of wellbeing and lived experience that take into account diversity on a local and global scale. Through interviews and observations, the research combined culturally relevant measures of emotional experience with responses to standardized life evaluations, such as the Cantrell ladder question used by the World Happiness Report, to contextualize the results.

[3] **Title**: “A Measure of Well-Being Efficiency Based on The World Happiness Report"
Authors: Francesco Sarracino, Kelsey O'Connor
Summary: They calculate an efficiency metric for well-being that evaluates how successfully a nation can convert inputs into subjective well-being. They

use Data Envelopment Analysis to a sample of 126 nations using the six inputs (real GDP per capita, healthy life expectancy, social support, freedom of choice, absence of corruption, and generosity) found in the World Happiness Reports. Efficiency scores show that countries with high subjective well-being, like the Nordics, are not necessarily the most efficient. The scores and economic efficiency have no correlation.

[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: They estimate a well-being efficiency metric that evaluates how well nations convert inputs into subjective well-being. Using Data Envelopment Analysis, we apply the six inputs (real GDP per capita, healthy life expectancy, social support, freedom of choice, lack of corruption, and generosity) found in the World Happiness Reports to a sample of 126 nations. Efficiency scores show that nations like the Nordics, which rank highly on subjective well-being, are not necessarily the most efficient. There is no correlation between the scores and economic efficiency.

[5] **Title**: "World Happiness Report (2012)"
Authors: John Helliwell, Richard Layard and Jeffrey Sachs
Summary: The Earth Institute produced the report, which Jeffrey Sachs, the institute's director, co-edited. It represents a growing global need for more emphasis on pleasure and the absence of misery as standards for public policy. It examines the current situation of happiness worldwide and demonstrates how regional and individual variances in happiness are explained by the emerging field of happiness science.

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 (happiness score), GDP, social support score, healthy and life expectancy, freedom of making life choices, kindness score, views regarding corruption, remaining from sources like Kaggle, WHO etc., to examine elements that each nation's population finds happy.

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:

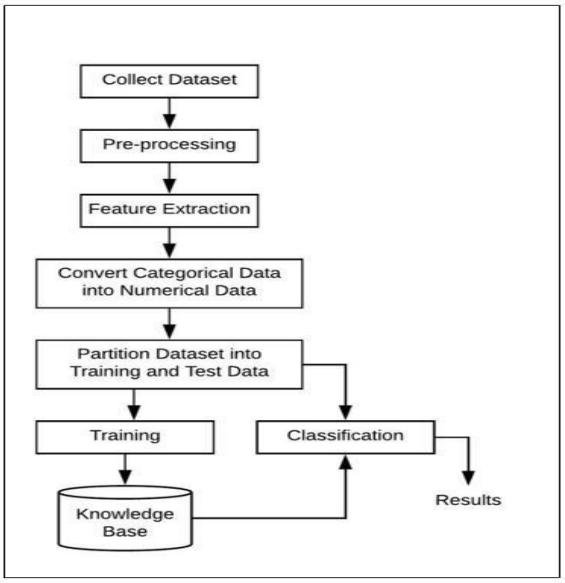

Fig. 1. Block Diagram

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.

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.

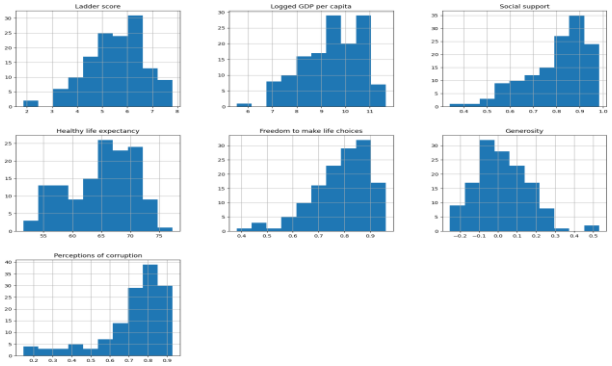

Fig. 2. Univariate analysis visualizations (Histogram)

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.

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

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.

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.

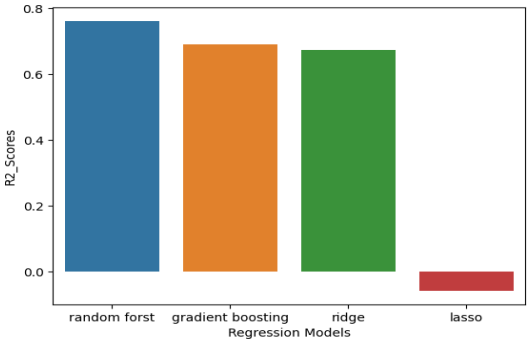

Fig. 4. R2score on test data of various models

VI. 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.