MATHEMATICAL MODELING AND COMPUTING, Vol. 10, No. 2, pp. 534–546 (2023)

Machine learning for the analysis of quality of life using the World Happiness Index and Human Development Indicators

Jannani A.*, Sael N., Benabbou F.

Laboratory of Information Technology and Modeling,
Faculty of sciences Ben M'Sik, Hassan II University of Casablanca,
Casablanca, Morocco
**Corresponding author: jannaniayoub@gmail.com*

(Received 31 January 2023; Accepted 22 April 2023)

Machine learning algorithms play an important role in analyzing complex data in research across various fields. In this paper, we employ multiple regression algorithms and statistical techniques to investigate the relationship between objective and subjective quality of life indicators and reveal the key factors affecting happiness at the international level based on data from the Human Development Index and the World Happiness Index covering the period from 2015 to 2021. The Pearson correlation analysis showed that happiness is related to the HDI score and GNI per capita. The best-performing model for forecasting happiness was the random forest regression, with a R2 score of 0.93667, a mean squared error of 0.0033048, and a root mean squared error of 0.05748, followed by the XGBoost regression and the Decision Tree regression, respectively. These models indicated that GNI per capita is the most significant feature in predicting happiness.

Keywords: *machine learning; quality of life indicators; Random Forest Regression; XGBoost Regression; Decision Tree Regression; statistical analysis.*

2010 MSC: 62H20, 62J05, 68T99 **DOI:** 10.23939/mmc2023.02.534

1. Introduction

Machine learning algorithms have gained widespread use in addressing social science issues [1, 2]. The advancement of modern society has led to the emergence of new concepts such as quality of life, which can benefit from the application of artificial intelligence models. These models can be used to analyze and understand the complex factors that contribute to QoL indicators, and can help in the development of policies and interventions aimed at improving wellbeing.

The World Health Organization defines quality of life as an individual's evaluation of their life situation within the cultural and value systems they exist in and in relation to their aspirations, norms, expectations, and worries [3]. This concept can be measured through both objective and subjective indicators, additionally, the latter are used more frequently [4]. Objective indicators include factors such as economy, health, education, and housing, while subjective indicators encompass personal experiences such as satisfaction and happiness in addition to people's perceptions of their health, life, and standard of living. Both objective and subjective indicators employ different data sources, each with its own set of advantages and limitations.

In this work, we aim to investigate the relationship between objective and subjective measures of QoL. In this context, the assessment of subjective wellbeing is based mainly on survey responses concerning the levels of satisfaction with standards of living. Therefore, we collected data from the World Happiness Report (WHR), which is published annually by the United Nations (UN), ranks nations according to their level of happiness based on global survey data from the Gallup World Poll (GWP) [5]. In our work, we consider the WHI's happiness score as a subjective wellbeing indicator. On the other hand, in order to examine objective indicators of well-being at the national level, we utilized the Human Development Index (HDI), which is produced by the UN in its Human Development Report. The HDI is based on three aspects of human development: a long and healthy life, knowledge,

and a decent standard of living [6]. The normalized indices for each of three dimensions' geometric means make up the HDI. The life expectancy at birth is used to evaluate the health dimension, while the mean number of years spent in school for persons 25 years of age and older and the anticipated number of years spent in school for young children are used to evaluate the knowledge dimension. GNI per capita is used to quantify the standard of living dimension [7, 8]. The remainder of this article is structured as follows: in the next section, we present related works to our issue, and in the third section, we present our methodology, including information about the datasets, statistical tools, and machine learning algorithms used. In the fourth section, we present and discuss the findings of our study. Finally, we provide conclusions and future research avenues.

2. Related work

In this section, we provide an overview of the research that has been conducted utilizing machine learning and statistical methods to study various country-level indicators of life quality, including the World Happiness Index and the Human Development Index. We present these works in chronological order.

The research [9] examines both physical and mental needs, including education, as factors that impact a nation's overall happiness. Over 90 features have been identified as predictors of a country's happiness. Due to the high number of features, manual analysis is not feasible; thus, support vector machines (SVM) and the information gain dimensionality reduction technique are used to predict happiness. Additionally, the SVM model achieved an accuracy of almost 90%. Data from 187 countries, sourced from the UN Development Project, is used to identify factors that need improvement to increase citizens' happiness. The study [10] examines the HDI and its components alongside the World Happiness Index in India and its neighboring South Asian nations. This work employed Pearson correlation to investigate the relationship between human development and happiness. Their results demonstrated a negative correlation, implying that happiness must be considered within the context of a particular cultural context. In [11], researchers used the Myers–Briggs personality type theory to analyze numerous national indicators, including the HDI, GDP per capita, and democracy index. Predictive models were built for each of the studied QoL indicators using four classification algorithms: Naive Bayes (NB), Multi-Layer Perceptron (MLP), SVM, and RF, and the best accuracy achieved was 72.8%. In [12], researchers analyzed the WHI 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 WHR. The research [13] developed a happiness prediction model using machine learning algorithms 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. The study employs techniques to address imbalanced data and achieves an overall prediction accuracy of 88.19% using DT. While the work [14] focused on using the World Happiness Report 2020 dataset to analyze the happiness index in Southeast Asia, the research employed exploratory data analysis (EDA) techniques to explore, analyze, and visualize the data. The results of the EDA are used to provide insights and recommendations for further data science processes in order to build data models. Researchers in [15] applied predictive modeling and Bayesian networks to model historical happiness index data of the WHR from 156 nations using multiple artificial intelligence algorithms, including general regression neural networks (GRNN), deep neural networks, RF, XGBoost, DT, ordinary least squares (OLS), and ridge regression, for prediction. The GRNN model achieved the best R-squared of 0.88, MAE of 0.29, and MSE of 0.15. The study also analyzed causal links among key factors impacting world happiness to provide useful information for policymaking. Additionally, [16] used regression analysis and correlation to investigate data from 150 countries on four indicators, including the World Happiness Index. The authors of [17] employed a variety of machine learning methods, including the Least Absolute Shrinkage and Selection Operator (LASSO), Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forests (RF), XGBoost, AdaBoost, and MLP, to forecast the WHI over data from the United Nations Sustainable Development Solutions Network (UN-

SDSN). The LASSO regressor was the best-performing model, with a R squared of 0.8954 and a RMSE of 0.0656. While the work [18] studied the relationship between global latitude and QoL measurements of the HDI and WHI using regression analysis and correlation. The authors of [19], combined two datasets, namely the happiness report dataset and the COVID-19 dataset, which comprises confirmed cases in multiple countries, to employ multiple machine learning algorithms, including linear regression, KNN, NB, SVM, and logistics regression. And in [20], authors conducted a study on WHR data to identify predictors of the WHI's happiness score using its other dimensions and used a variety of algorithms, which are: MLP, RF, One Rule, and XGBoost. The RF model achieved the best accuracy of 92.4%.

This literature review shows that RF, MLP, and SVM/SVR are the most commonly utilized algorithms for studying the WHI and HDI, in addition to NB and XGBoost. These algorithms are used to build predictive classification and regression models to satisfy a variety of research goals. Regression analysis, correlation techniques, and EDA are also frequently used in these studies for analysis.

In our research, we utilize exploratory data analysis and statistical techniques, such as Pearson correlation and principal component analysis, to examine the associations within our data. Additionally, we utilize machine learning methods to further investigate the effect of objective factors on the happiness score of the World Happiness Index.

3. Research methodology

The current study aims to study the relationships between various wellbeing indicators and their potential impact on happiness. A comprehensive analysis of the data is conducted using both statistical techniques and machine learning algorithms to identify human development indicators that are key predictors of happiness as measured by the World Happiness Index. Figure 1 illustrates the key procedures involved in the experiment. It provides a visual representation of the different stages of the experiment, starting with the preparation of the data, followed by the actual experimental process, and finally the analysis of the results. The figure depicts the process of the experiment and its various steps, facilitating easy comprehension of the experiment's methodology.

3.1. Data collection

This study is based on a thorough examination of two datasets: the first, the Human Development Index measures [21], procured from the Data Center for the Human Development Report. This dataset encompasses a variety of indicators, including: the HDI scores, GNI per capita, inequality in income, the gender development index, the gender inequality index, life expectancy at birth, expected years of schooling, mean years of schooling, inequality in education, the inequality-adjusted HDI, inequality in life expectancy, the maternal mortality ratio, the adolescent birth rate, male and female labor force participation rates, the planetary pressures-adjusted HDI, carbon dioxide emissions per capita, and material footprint per capita. Additionally, this study also includes data from The World Happiness Report [5], which ranks countries based on six features: GDP per capita, freedom, social support, government corruption, life expectancy, and generosity. These datasets provide a comprehensive understanding of the subject matter being examined and serve as the foundation for the research conducted in this study.

3.2. Data preparation

As our goal is to explore correlations between attributes rather than the well-being of particular countries, we merged the data and eliminated any null values. This led to data pertaining to 119 nations for the seven-year interval spanning from 2015 to 2021. Afterwards, the features were standardized using the Standard Scaler technique, which is a procedure employed to normalize numerical data through the elimination of the mean and transformation to unit variance [22, 23]. Standardizing data can be useful for many algorithms that work better with normally distributed data and can also be required for improving the performance of some algorithms that are sensitive to the scale of the input features.

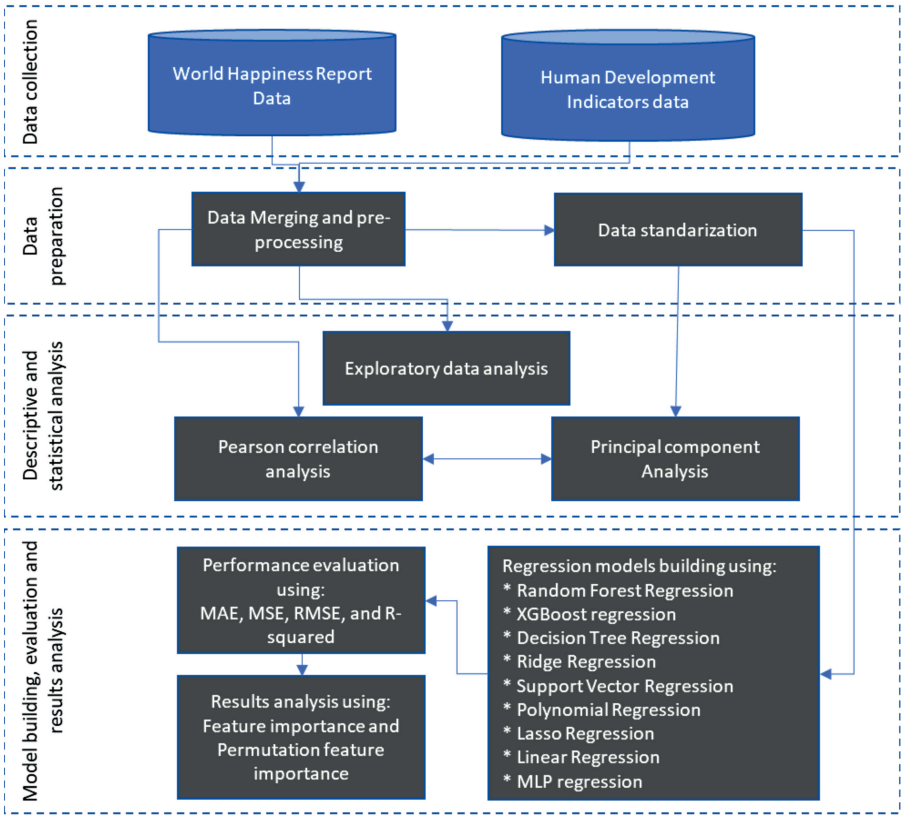

Fig. 1. Steps of the proposed experimentation.

3.3. Descriptive and statistical analysis

A comprehensive descriptive and statistical analysis is performed to examine the distribution and variation of the overall HDI and the happiness score of the World Happiness Index. To accomplish this, two types of charts were utilized: a line chart to investigate the temporal relationship between the measures and a scatter plot to analyze the distribution of these measures across geographical regions. Additionally, to carry out statistical analysis, we utilized Pearson correlation and principal component analysis. These methods were employed to uncover any inter-relationships and patterns within the data, providing further insights and understanding into the distribution of the data.

3.4. Model building, evaluation, and results analysis

In the subsequent phase of the study, a variety of regression algorithms are employed on the data to determine the most significant indicators of the HDI as predictors of happiness. The algorithms utilized include RF regression, XGBoost regression, DT regression, LASSO regression, ridge regression, SVR, polynomial regression, linear regression, and MLP Regression. These models were subsequently evaluated utilizing four performance metrics: R-squared, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). The proposed models are then analyzed and interpreted to identify the crucial predictors of happiness based on the built-in feature importance and permutation feature importance of the top three models with optimal performance.

4. Experiment and results

4.1. Exploratory data analysis

In order to extract valuable information from the data, we employed the programming language Python during the initial stages of data cleansing and preparation. Subsequently, the cleaned data was then input into Tableau in order to analyze and visualize our data. Our primary focus of study was centered around the HDI and WHI scores, as depicted in Figure 2, which illustrates the distribution of these scores by region.

Upon conducting our analysis, we observed a linear distribution of the HDI and WHI scores, which confirms the existence of a relationship between these two measures. Furthermore, we noted a correlation between the wellbeing of countries belonging to specific geographical regions. As an example, countries located in Western Europe consistently ranked at the top of both measures, while nations situated in the sub-Saharan region recorded the lowest scores. This highlights the significance of geographic location in determining the overall wellbeing of a country. The cultural contexts are important perspectives to evaluate happiness [10], which correlates with the projection of QoL scores by geographical regions as we present below.

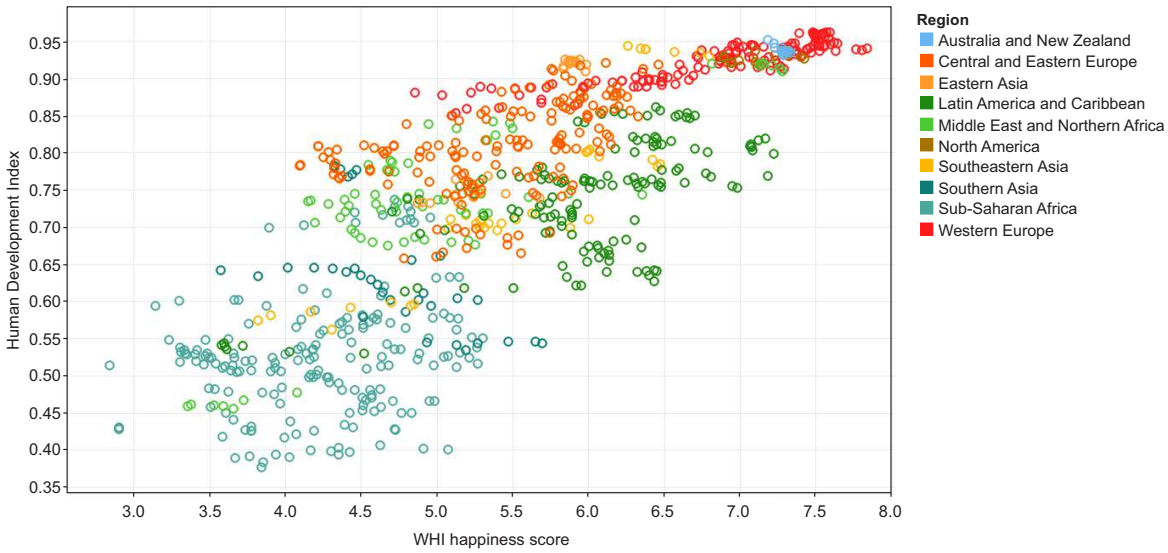

Fig. 2. Scatter plot of WHI and HDI by region.

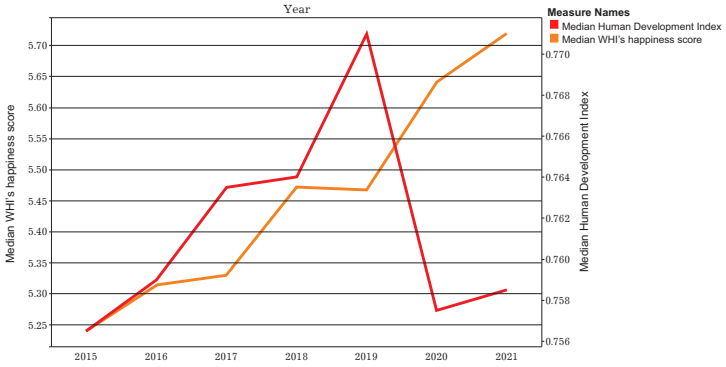

Fig. 3. Line chart of median values of HDI and WHI by year.

The line chart presented in Figure 3 displays the median values of the HDI and the WHI scores, represented by years. The chart illustrates the change in median values over time, allowing for an analysis of the trend and patterns of these indices. The line chart is a useful tool for comparing the relative scores of the indexes and identifying any significant changes or fluctuations in the median values. The chart is also useful in identifying and determining any correlation or relationship between two indices.

This figure illustrates that prior to the year 2019, the fluctuations of the HDI and WHI scores were relatively similar. However, after this year, we observed a deviation in the trend where the yearly median HDI scores decreased while the WHI scores continued to increase. This discrepancy in the trend between these two indexes may be indicative of a shift in the factors that impact them. The COVID-19 pandemic is likely to be the underlying cause of the observed decline in the mean HDI scores in the year 2020. However, it appears that the pandemic did not have a significant impact on the WHI scores, which continued to rise. According to the World Happiness Report 2020, the increase in wellbeing scores in the Nordic countries, for instance, is attributed to the social and environmental dimensions. One of the key determinants of wellbeing in these countries, especially during the pandemic, is the level of trust at both institutional and personal levels [24]. This highlights the importance of social factors in determining the overall wellbeing of a country, especially during the pandemic.

4.2. Correlation analysis and principal component analysis

Upon conducting the exploratory descriptive analysis, the next step in this study is to perform a statistical analysis to identify the features that significantly influence our data. To accomplish this task, we used Pearson correlation and principal component analysis methods.

The correlation between two data points X and Y is quantified by the degree of linear relationship between their respective attributes. The Pearson Correlation is a metric that ranges from -1 to $+1$, representing a perfect negative and positive correlation, respectively. A value of zero indicates the absence of correlation between the studied variables [25].

We employ Pearson correlation to investigate the relationship between the various dimensions of QoL of both the HDI and WHI indicators and their sub-indicators. As depicted in Figure 4, a substantial portion of the WHI indicators displayed a strong correlation with a majority of the HDI indicators, particularly the GDP per capita, the happiness score, life expectancy, and social support.

The happiness score exhibited a positive correlation with the HDI score, the inequality-adjusted HDI, and GNI per capita, and displayed a negative correlation with the Gender Inequality Index, inequality in life expectancy, and inequality in education. Additionally, the HDI score displayed a significant positive correlation with GDP per capita, life expectancy, and the happiness score. The correlation between the WHI rank and both the happiness and HDI scores is negative; this is due to the fact that countries with the highest QoL scores have the lowest ranking values (top values).

	WHI corruption perception	WHI freedom	WHI gdp per capita	WHI generosity	WHI happiness rank	WHI happiness score	WHI life expectancy	WHI social support
Human Development Index	0,023284296	0,235542708	0,952418304	-0,006968645	-0,80821845	0,808681085	0,879506233	0,560162553
Life Expectancy at Birth	-0,017986519	0,205203227	0,873761456	0,031493457	-0,776070503	0,769250599	0,933151204	0,527776148
Expected Years of Schooling	0,039697512	0,242088745	0,836508373	-0,001144248	-0,742693015	0,745071935	0,783995224	0,483500132
Mean Years of Schooling	0,042954078	0,208520854	0,860848031	-0,029261481	-0,69374901	0,699467355	0,764972804	0,519826649
Gross National Income Per Capita	0,040232686	0,272230894	0,861075771	0,113459213	-0,784582731	0,788881034	0,732935927	0,490185316
Gender Development Index	0,025741402	0,228595492	0,534372471	-0,028229955	-0,518210172	0,507303977	0,556725667	0,420760955
Inequality-adjusted Human Development Index	0,02925737	0,234470584	0,926230008	-0,000204104	-0,787494602	0,789811797	0,874275831	0,556631646
Inequality in life expectancy	-0,053479003	-0,22761442	-0,847284628	0,076651183	0,711057021	-0,706719578	-0,900734736	-0,506425487
Inequality in eduation	-0,01927725	-0,221413913	-0,776494642	0,010272668	0,662189505	-0,658506009	-0,743366758	-0,53798971
Inequality in income	-0,024464424	-0,045712018	-0,306733504	-0,021746755	0,227625676	-0,248889452	-0,336479349	-0,16019209
Gender Inequality Index	-0,046420661	-0,23883234	-0,854022989	0,010740729	0,748137653	-0,748300308	-0,838572804	-0,510677419
Maternal Mortality Ratio (deaths per 100,000 live births)	-0,013758338	-0,184439787	-0,723850088	0,080862591	0,59360667	-0,585541163	-0,801832205	-0,421937384
Adolescent Birth Rate (births per 1,000 women ages 15-19)	-0,061642328	-0,175678638	-0,785957562	0,009984624	0,596609175	-0,596841627	-0,783086994	-0,387247342
Labour force participation rate, female (% ages 15 and older)	-0,044311427	0,13953442	-0,155384912	0,236849472	-0,02445355	0,012043149	-0,103395753	0,031410215
Labour force participation rate, male (% ages 15 and older)	-0,082351752	0,052145328	-0,330125815	0,205566752	0,158816646	-0,175032514	-0,192288461	-0,117990256
Planetary pressures-adjusted Human Development Index	0,019889221	0,176377848	0,806594476	-0,127751195	-0,656547755	0,65150642	0,800662688	0,463886254
Carbon dioxide emissions per capita (production) (tonnes)	0,005116697	0,108582721	0,703019525	0,107859147	-0,543642543	0,552090045	0,548557516	0,438621578
Material footprint per capita (tonnes)	0,021827303	0,246023906	0,769108973	0,135023008	-0,71112062	0,716526349	0,647010155	0,461535856

Fig. 4. Pearson correlation between WHI and HDI indicators.

Principal component analysis (PCA) is a statistical method used to simplify large datasets by reducing their dimensionality while retaining important information. This is achieved by creating new, uncorrelated variables that optimize the variance in the data [26]. However, in our work, we are interested in its explained variance, which represents the extent to which the variation in our dataset can be statistically explained by each of the principal components. We use it as a metric to select the principal components that are most relevant to our analysis. As depicted in Figure 5, it can be observed that the top three components account for more than 80% of the total variance. Consequently, it was deemed appropriate to retain these three components for further analysis in the subsequent steps. These components are considered to be of significant importance and relevance to our analysis and will be used to identify the crucial factors that exert the greatest influence on our data.

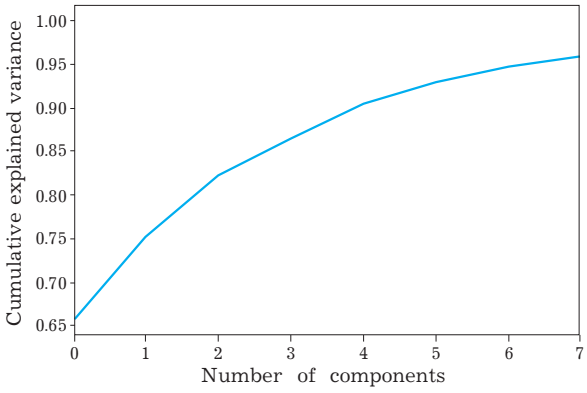

Fig. 5. The scree plot of principal components by cumulative explained variance.

	Component1	Component2	Component3
Component1	1,0000	0,0000	0,0000
Component2	0,0000	1,0000	0,0000
Component3	0,0000	0,0000	1,0000
Human Development Index	-0,9872	0,0046	-0,0335
Life Expectancy at Birth	-0,9322	0,0334	-0,0623
Expected Years of Schooling	-0,9052	-0,0801	-0,0097
Mean Years of Schooling	-0,9318	0,0103	-0,0466
Gross National Income Per Capita	-0,8344	-0,1409	0,3372
Gender Development Index	-0,6649	-0,4487	-0,3432
Inequality-adjusted Human Development Index	-0,9922	-0,0051	0,0395
Inequality in life expectancy	0,9459	-0,0323	0,1868
Inequality in eduation	0,8906	0,1284	0,1588
Inequality in income	0,4035	-0,1988	-0,5695
Gender Inequality Index	0,9539	0,0679	-0,0927
Maternal Mortality Ratio (deaths per 100,000 live births)	0,8209	-0,1303	0,3218
Adolescent Birth Rate (births per 1,000 women ages 15-19)	0,8793	-0,2198	-0,0251
Labour force participation rate, female (% ages 15 and older)	0,0636	-0,9207	0,1209
Labour force participation rate, male (% ages 15 and older)	0,3454	-0,6691	-0,1524
Planetary pressures-adjusted Human Development Index	-0,8564	0,1585	-0,4001
Carbon dioxide emissions per capita (production) (tonnes)	-0,7093	-0,0555	0,3658
Material footprint per capita (tonnes)	-0,7640	-0,2219	0,4334
WHI's happiness score	-0,7837	-0,1168	0,0711

Fig. 6. Pearson correlation of selected principal components with HDI indicators and WHI happiness score.

Figure 6 illustrates the correlation between three selected principal components, the HDI indicators, and the WHI's happiness score. It is noteworthy that PCA is a technique that eliminates collinearity; thus, five components in this figure do not display any correlations with one another.

Furthermore, the chosen components contain the most significant information, with the first component in particular containing nearly 75% of the accumulated explained variance. It can be observed that this component exhibits the strongest positive correlation with the Gender Inequality Index, inequality in life expectancy, and inequality in education, and the strongest negative correlation with the inequality adjusted HDI, the HDI score, mean years of schooling, and life expectancy at birth. Further, the second component demonstrates a correlation with the labor force participation indicators and gender development index, while the other components have no significant correlations with the features.

4.3. Model building

Regression analysis is a widely-used statistical method that allows for the quantification of the degree of association between variables and the estimation of the impact of one variable on another [27]. In this study, our objective is to conduct a regression analysis to investigate the relationship between the subjective measurement of wellbeing, represented by the WHI happiness score, and the objective indicators of the HDI. The goal of this analysis is to determine the extent to which the HDI's metrics are associated with the happiness score and to estimate the impact of the objective metrics on a subjective measurement of wellbeing. This analysis will provide valuable insights for policy goals and help to understand how different aspects of human development are related to overall happiness and well-being.

Used algorithms. After conducting the review of literature, as we presented in the related work section, it was determined that certain algorithms are commonly employed for the examination of well-being indicators on an international level. In light of this, the identified algorithms were implemented in conjunction with additional algorithms in order to build predictive regression models, evaluate their respective performances, and determine the most effective ones. In this study, a diversity of regression models are proposed, which are:

1. Linear Regression: A method that endeavors to establish a linear relationship between two variables by applying a linear equation [28].
2. LASSO Regression: An algorithm that determines the linear association between input data and targets so that correlated features are excluded during model building to improve the model and prevent overfitting [29].
3. Ridge Regression: a method of linear regression that incorporates the L2 regularization aiming is to minimize both the sum of squared residuals and the L2 norm of the coefficients, resulting in their shrinking towards zero. This serves to counteract overfitting and improve the model's ability to be applied to new data [30, 31].

4. Polynomial Regression: The method of polynomial regression serves as a comprehensive approach for building upon the linear regression technique in order to effectively fit curved lines to the response data [32]. In this work, we use the second-degree polynomial regression.
5. SVR: It is a regression algorithm that employs the same principle as SVM. It calculates a linear regression function within a high-dimensional feature space, where input data is transformed through a nonlinear function [33].
6. XGBoost regression: XGBoost is a boosting model designed for ensemble learning that endeavors to attain exceptional performance through the utilization of weak classifiers, specifically decision trees, in its prediction tasks [34].
7. Decision Tree Regression: It is a supervised algorithm applied to solve multiple regression problems. The regression trees approximate an unknown regression function using a tree-based method, with a sample dataset as input. The models generated include a hierarchy of logical conditions based on the predictor variables. The final nodes, referred to as "leaves", hold the numerical predictions for the target variable [35].
8. Random Forest Regression (RFR): it is an algorithm that generates predictions by aggregating the outcomes of a sequence of regression decision trees [36].
9. MLP Regression: a type of neural network composed of three essential layers: the input layer, which receives the data; the output layer, accountable for executing the prediction task; and one or multiple hidden layers serving as the central computational component of the MLP [37].

Performance metrics. To assess the efficacy of regression models, two aspects must be studied: the error and the goodness of fit [38]. In our work, we use three metrics for assessing error, namely mean absolute error, root mean squared error, and mean squared error, in conjunction with the R squared coefficient of determination.

4.4. Results

Performance evaluation. The results of our experiment are outlined in the table below. Our analysis reveals that the Random Forest Regressor algorithm exhibits the highest level of performance among the algorithms utilized, as evidenced by its R2 score of 0.93667, MSE of 0.0033048, and RMSE of 0.05748. It is worth noting that the XGBoost regressor also displayed comparable results to the RFR and even outperformed it in terms of the MAE, which was recorded at 0.0377065. Lastly, the decision tree regression algorithm placed third in the table, with an R2 score surpassing 0.90. These three models will be used for further analysis in the next section.

Table 1. Comparative table of the proposed regression models.

Model	Performance			
	MAE	MSE	RMSE	R2 Score
Random Forest Regression	0.0401992	0.0033048	0.05748	0.93667
XGBoost regression	0.0377065	0.0035850	0.05987	0.931300
DT Regression	0.0466717	0.0050338	0.07095	0.903538
Ridge Regression	0.0562938	0.0062312	0.07893	0.880591
SVR	0.0586839	0.0064068	0.08004	0.877227
Polynomial Regression	0.0786452	0.0115333	0.10739	0.77899
Lasso Regression	0.0881330	0.0123623	0.11118	0.763103
Linear Regression	0.0940887	0.0132159	0.11496	0.746746
MLP regression	0.0950311	0.0134265	0.11587	0.742710

Feature importance of the best performing models. There are several techniques for determining the importance of features within a dataset, and the most suitable technique depends on the context and nature of the data. In the current study, we use the default feature importance attribute of three top-performing models in conjunction with the technique of permutation importance.

The feature importance attribute produces an array of values that signify the significance of each feature within the dataset. These values are computed through techniques specific to the model, such

as the mean decrease impurity method for decision trees or the mean decrease accuracy method for random forests [39, 40]. In the case of XGBoost, the importance score is determined by averaging the gain across all splits that include the feature, where gain is a measure of the feature's contribution to reducing the loss function. Higher values of gain correspond to more important features [41].

Table 2 presents the outcome of the examination, wherein the highest score among the utilized features is for GNI per capita across all the examined models, with a score of 0.776 for the decision tree regressor, 0.710 for the random forest regressor, and 0.645 for the XGBoost regressor. The other features obtained relatively low scores in comparison to the top-ranking feature and were ranked differently for each model. The top five dimensions for the random forest regressor, in addition to GNI per capita, are the HDI score, income inequality, gender inequality index, and maternal mortality ratio. Conversely, for the XGBoost regression model, the maternal mortality ratio, the HDI score, income inequality, and labor force participation rate are in the top five, in that order. Furthermore, for the Decision Tree Regressor model, the labor force participation rate of males, income inequality, and adolescent birth rate followed GNI per capita.

To analyze the results of our models, we have compiled the results of each feature from three models. The overall total of the importance scores indicates that the highest-ranked factors include GNI per capita, the ratio of maternal mortality, inequality in income, the HDI score, and the rate of male participation in the labor force.

Table 2. The importance of each feature in relation to the three selected predictive models.

Features	Feature importance for each model			
	Random Forests regressor	XGBoost regressor	Decision Tree regressor	Total
Gross National Income Per Capita	0.710	0.645	0.776	2.132
Maternal Mortality Ratio	0.023	0.074	0.001	0.097
Inequality in income	0.036	0.035	0.024	0.094
Human Development Index	0.039	0.042	0.004	0.085
Male labour force participation rate	0.021	0.026	0.033	0.080
Adolescent Birth Rate	0.019	0.024	0.024	0.067
Gender Inequality Index	0.027	0.016	0.017	0.059
Material footprint per capita	0.009	0.024	0.021	0.054
Inequality in life expectancy	0.013	0.016	0.020	0.050
Female labour force participation rate	0.016	0.011	0.016	0.042
Gender Development Index	0.012	0.010	0.015	0.037
Carbon dioxide emissions per capita	0.016	0.011	0.009	0.036
Inequality in eduation	0.009	0.012	0.012	0.033
Inequality-adjusted HDI	0.008	0.008	0.004	0.021
Expected Years of Schooling	0.009	0.007	0.004	0.020
Mean Years of Schooling	0.009	0.005	0.004	0.017
Life Expectancy at Birth	0.009	0.003	0.002	0.014
Planetary pressures-adjusted HDI	0.005	0.003	0.003	0.011

Permutation importance is a method for determining the significance of a feature in a predictive model by evaluating the alteration in its accuracy when the values of that feature are randomly permuted [42].

Table 3 shows the results of this measure, wherein the importance scores for each feature are presented in accordance with the three models. Additionally, an overall total feature importance measure is also provided. Upon examination of the overall measure scores, it can be deduced that the GNI per capita is the most significant feature of the model, which aligns with the findings previously presented in Table 2. The subsequent most important features are the HDI score, income inequality, adolescent birth rate, and labor force participation rate.

In conclusion, it was determined that the results obtained from the use of both techniques were similar, with the exception of the maternal mortality ratio feature. The remaining features were determined to be among the top five features, and it was determined that the feature of GNI per capita had the most significant impact among the features that were studied.

Table 3. The permutation importance of for the three selected models.

Features	Feature permutation importance scores for each model			
	Random Forests regressor	XGBoost regressor	Decision Tree regressor	Total
Gross National Income Per Capita	1.3714639	0.9783138	1.6581317	4.0079094
Human Development Index	0.0376951	0.1400567	0.0278490	0.2056008
Inequality in income	0.0574029	0.0259451	0.0963287	0.1796767
Adolescent Birth Rate	0.0175036	0.0577834	0.0856210	0.1609080
Male labour force participation rate	0.0201728	0.0356807	0.0802950	0.1361484
Female labour force participation rate	0.0236162	0.0332104	0.0645155	0.1213421
Gender Inequality Index	0.0318550	0.0245717	0.0559147	0.1123414
Gender Development Index	0.0131438	0.0048898	0.0939651	0.1119988
Inequality in life expectancy	0.0125592	0.0170812	0.0365263	0.0661668
Inequality in eduation	0.0139042	0.0258071	0.0180006	0.0577119
Carbon dioxide emissions per capita	0.0105202	0.0088934	0.0357508	0.0551643
Inequality-adjusted HDI	0.0116216	0.0078699	0.0262755	0.0457671
Material footprint per capita	0.0054602	0.0123214	0.0249009	0.0426825
Life Expectancy at Birth	0.0078995	0.0124033	0.0152358	0.0355385
Maternal Mortality Ratio	0.0081633	0.0256182	0.0016141	0.0353956
Mean Years of Schooling	0.0089043	0.0119298	0.0036324	0.0244665
Expected Years of Schooling	0.0084232	0.0073736	0.0085562	0.0243530
Planetary pressures-adjusted HDI	0.0023403	-0.0004786	0.0063412	0.0082029

5. Conclusion

In this article, we delved deeper into the correlation between various indicators of well-being at the country level. Through our analysis of the relationship between indicators from the World Happiness Index and the Human Development Index, we sought to understand the key factors that are significant in determining happiness.

The results, obtained through the use of Pearson correlation analysis, identified a strong correlation between the overall measures of the HDI and the happiness score. This indicates that countries with higher HDI scores tend to have higher levels of happiness as well. Furthermore, through the use of PCA, we found that the first component correlated with a variety of indicators. In addition, our correlation analysis also revealed that multiple indicators explained a high variance and therefore contained more information about the concept of well-being. This highlights the complexity of measuring well-being and the importance of considering multiple indicators to gain a comprehensive understanding.

Our regression models further demonstrated that gross national income per capita, the HDI, inequality in income, and the male labor force participation rate were significant predictors of happiness, which shows that at least three of the top five features represent economic dimensions. This confirms the results of the Pearson correlation and highlights the importance of economic indicators in determining levels of happiness.

In future research, we plan to expand our study to include more indicators of well-being and further investigate their relationships, particularly with regard to other economic aspects of well-being and their impact on happiness. This will allow us to gain a more comprehensive understanding of the determinants of well-being and inform policy-making aimed at promoting happiness and well-being.

[1] Radford J., Joseph K. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. Frontiers in Big Data. **3**, 18 (2020).

[2] Grimmer J., Roberts M. E., Stewart B. M. Machine Learning for Social Science: An Agnostic Approach. Annual Review of Political Science. **24** (1), 395–419 (2021).

[3] WHOQOL – Measuring Quality of Life| The World Health Organization. https://www.who.int/tools/whoqol.

[4] Davis E., Waters E., Shelly A., Gold L. Children and Adolescents, Measuring the Quality of Life of. International Encyclopedia of Public Health. 641–648 (2008).

[5] Helliwell J. F., Layard R., Sachs J. D., Neve J.-E. D., Aknin L. B., Wang S. World Happiness Report (2022). https://worldhappiness.report/ed/2022/.

[6] Human Development Index, United Nations. https://hdr.undp.org/data-center/human-development-index.

[7] Taner M., Sezen B., Mihci H. An Alternative Human Development Index Considering Unemployment. South East European Journal of Economics and Business. **6** (1), 45–60 (2011).

[8] Martinez R. Inequality and the new human development index. Applied Economics Letters. **19** (6), 533–535 (2012).

[9] 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).

[10] Basu R., Behera S. K., Adak D. K. Human Development and Happiness: Are the Two Interlinked? International Journal of Indian Psychology. **6** (3), 141–150 (2018).

[11] 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 Intelligence, Robotics & Optimization (ICCAIRO). 48–52 (2018).

[12] Carlsen L. Happiness as a sustainability factor. The world happiness index: a posetic-based data analysis. Sustainability Science. **13** (2), 549–571 (2018).

[13] 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).

[14] Riyantoko P. A. Southeast Asia Happiness Report in 2020 Using Exploratory Data Analysis. International Journal of Computer, Network Security and Information System. **2** (1), 1 (2020).

[15] Dixit S., Chaudhary M., Sahni N. Network Learning Approaches to study World Happiness. ArXiv:2007.09181 (2020).

[16] Okagbue H. I., Oguntunde P. E., Bishop S. A., Adamu P. I., Akhmetshin E. M., Iroham C. O. Significant Predictors of Henley Passport Index. Journal of International Migration and Integration. **22** (1), 21–32 (2021).

[17] Jannani A., Sael N., Benabbou A. Predicting Quality of Life using Machine Learning: case of World Happiness Index. 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). 1–6 (2021).

[18] Pawliczek A., Kurowska-Pysz J., Smilnak R. Relation between Globe Latitude and the Quality of Life: Insights for Public Policy Management. Sustainability. **14** (3), 1461 (2022).

[19] Farooq S. A., Shanmugam S. K. A Performance Analysis of Supervised Machine Learning Techniques for COVID-19 and Happiness Report Dataset. Sentimental Analysis and Deep Learning. 591–601 (2022).

[20] Khder M. A., Sayf M., Fujo S. W. Analysis of World Happiness Report Dataset Using Machine Learning Approaches. International Journal of Advances in Soft Computing and its Applications. **14** (1), 15–34 (2022).

[21] Home. Human Development Reports. https://hdr.undp.org/.

[22] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. **12** (85), 2825–2830 (2011).

[23] sklearn.preprocessing.StandardScaler scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[24] Helliwell J. F., Huang H., Wang S., Norton M. World Happiness, Trust and Deaths under COVID-19 (2021).

[25] Nettleton D. Chapter 6 – Selection of Variables and Factor Derivation. Commercial Data Mining. Processing, Analysis and Modeling for Predictive Analytics Projects. 79–104 (2014).

[26] Jolliffe I. T., Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. **374** (2065), 20150202 (2016).

[27] Angelini C. Regression Analysis. Reference Module in Life Sciences. Encyclopedia of Bioinformatics and Computational Biology. **1**, 722–730 (2019).

[28] Shobha G., Rangaswamy S. Chapter 8 – Machine Learning. Handbook of Statistics. **38**, 197–228 (2018).

[29] Misra S., Li H., He J. Chapter 5 – Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods. Machine Learning for Subsurface Characterization. 129–155 (2020).

[30] Fathi E., Shoja B. M. Chapter 9 – Deep Neural Networks for Natural Language Processing. Handbook of Statistics. **38**, 229–316 (2018).

[31] Simske S. Chapter 4 – Meta-analytic design patterns. Meta-Analytics. 147–185 (2019).

[32] Banks D. L., Fienberg S. E. Statistics, Multivariate. Encyclopedia of Physical Science and Technology (Third Edition). 851–889 (2003).

[33] Basak D., Pal S., Patranabis D. Support Vector Regression. Neural Information Processing – Letters and Reviews. **11** (10), 203–224 (2007).

[34] Dong J., Chen Y., Yao B., Zhang X., Zeng N. A neural network boosting regression model based on XGBoost. Applied Soft Computing. **125**, 109067 (2022).

[35] Torgo L. Regression Trees. Encyclopedia of Machine Learning and Data Mining. 1080–1083 (2017).

[36] Williams B., Halloin C., Löbel W., Finklea F., Lipke E., Zweigerdt R., Cremaschi S. Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction. Computer Aided Chemical Engineering. **48**, 1639–1644 (2020).

[37] Abirami S., Chitra P. Chapter Fourteen – Energy-efficient edge based real-time healthcare support system. Advances in Computers. **117** (1), 339–368 (2020).

[38] Chicco D., Warrens M. J., Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science. **7**, e623 (2021).

[39] Benard C., Da Veiga S., Scornet E. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. Biometrika. **109** (4), 881–900 (2022).

[40] Scornet E. Trees, forests, and impurity-based variable importance. ArXiv:2001.04295 (2020).

[41] Shi X., Wong Y. D., Li M. Z.-F., Palanisamy C., Chai C. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention. **129**, 170–179 (2019).

[42] 4.2. Permutation feature importance. https://scikit-learn/stable/modules/permutation_importance.html.

Машинне навчання для аналізу якості життя за допомогою світового індексу щастя та індикаторів людського розвитку

Джанані А., Сейл І., Бенаббу Ф.

Лабораторія інформаційних технологій та моделювання,
Факультет наук Бен М'Сік, Університет Хасана II Касабланки,
Касабланка, Марокко

Алгоритми машинного навчання відіграють важливу роль в аналізі складних даних у дослідженнях у різних сферах. У цій статті використовуються алгоритми множинної регресії та статистичні методи для дослідження зв'язку між об'єктивними та суб'єктивними показниками якості життя та виявлення ключових факторів, що впливають на щастя на міжнародному рівні на основі даних індексу людського розвитку та Світового індексу щастя, що охоплюють період з 2015 по 2021 рік. Кореляційний аналіз Пірсона показав, що щастя пов'язане з показником індексу людського розвитку та валовим національним доходом на душу населення. Найефективнішою моделлю для прогнозування щастя була випадкова лісова регресія з показником R2 0.93667, середньоквадратичною помилкою 0.0033048 і середньоквадратичним значенням 0.05748, за якою йшли регресія XGBoost і регресія дерева рішень відповідно. Ці моделі показали, що валовий національний дохід на душу населення є найважливішою характеристикою для прогнозування щастя.

Ключові слова: *машинне навчання; показники якості життя; регресія випадкового лісу; регресія XGBoost; регресія дерева рішень; статистичний аналіз.*