Relevant factors that intervene in world happiness, analysis of countries with high happiness and low GDP.

[2] Data Science Research Perú, Perú, milton.otiniano@hotmail.com

***Abstract*– The objective of this article is to achieve a statistical analysis of the data from the World Happiness Report and its relationship with the economic income of the countries, Gross Domestic Product (GDP), as well as to identify other variables that influence happiness. The approach uses research methodology and Machine Learning techniques to generate models representing the function to predict a country's happiness score or category, as well as a hypothesis test.**

The results determine that there is a positive relationship between economic income and happiness, but, up to a certain level, the relationship is marginal, and that there are other aspects, such as those related to social factors, which are decisive to establish the degree of happiness in a country.

Keywords-- happiness, GDP, country economy, Python, Machine learning.

Digital Object Identifier: http://dx.doi.org/10.18687/LACCEI2021.1.1.303
ISBN: 978-958-52071-8-9 **ISSN**: 2414-6390
DO NOT REMOVE

Relevant factors that intervene in world happiness, analysis of countries with high happiness and low GDP.

[2] Data Science Research Perú, Perú, milton.otiniano@hotmail.com

***Abstract*– The objective of this article is to achieve a statistical analysis of the data from the World Happiness Report and its relationship with the economic income of the countries, Gross Domestic Product (GDP), as well as to identify other variables that influence happiness. The approach uses research methodology and Machine Learning techniques to generate models representing the function to predict a country's happiness score or category, as well as a hypothesis test.**

The results determine that there is a positive relationship between economic income and happiness, but, up to a certain level, the relationship is marginal, and that there are other aspects, such as those related to social factors, which are decisive to establish the degree of happiness in a country.

Keywords-- happiness, GDP, country economy, Python, Machine learning.

I. INTRODUCTION

Happiness has always been the objective of human beings, from the time of Aristotle who mentioned "Happiness is the meaning and purpose of life, the general and final goal of human existence". Later, Maslow (1943) defined happiness as the self-fulfillment that individuals achieve after satisfying, partially or totally, certain hierarchically ordered needs [1]. The last resolution issued by the United Nations indicates that "the pursuit of happiness is a fundamental human goal and embodies the spirit of the globally agreed targets known as the Millennium Development Goals". [2]

In the 21st century, the field of social psychology has been expanding into other areas and happiness began to be studied from a social science approach. In comparison with those who feel they do not have other people they can trust, people who feel they have adequate social support tell they are happier, and it has also been found that they have fewer psychological problems, including eating disorders and mental illness. [3] [4]

Over the years, happiness has been analyzed from different perspectives, relating it to well-being and positive psychology. [5] It is also related to "the economics of happiness" that reports empirical associations between happiness and other variables [6].

On the other hand, with the arrival of Industry 4.0, studies began to be carried out at community or national level, since it is known about the existence of governance systems that are gradually deteriorating the well-being situation, because of the implementation of management models, based on job insecurity and the massive reduction in jobs derived from the automation of production processes and the extensive use of robots. All these factors negatively influence the happiness of human beings, especially in ecosystems that are far from the guiding principles of well-being and justice. [7]

Other studies were developed by starting from preconceived ideas, with a lot of content and wisdom that invites reflection, but without real evidence about the degree of happiness of people and what makes them happy, which is what the social scientist wants to discover. [8]

Evenly, happiness is increasingly attracting more attention - from politicians and decision-makers - both in developed and developing countries, in order to answer questions such as, what are the factors that influence the well-being of society to be included in the public policy guide? Is GDP the only factor?

Recent authors mention that happiness is a much more complete measure than GDP, since it refers to the economic aspect of life, focusing only on production and income. [9] A Harvard study, almost 80 years old, has proved that embracing community helps us live longer, and be happier. [10]

Since 2012, the World Happiness Report has been published [11]. This report shows the state of happiness of 156 participating countries. This index takes into account the following factors that could be determining for the development of towns:

1. GPD per capita
2. Social support
3. Healthy life expectancy
4. Freedom to make life choices
5. Generosity
6. Perception of corruption

To implement this analysis, the annual happiness report data was accessed from the Kaggle data repository and the 2019 report was consulted. [12]

II. RELATED RESEARCH

A. Research 1: Does income influence the happiness of populations? The cases of Colombia, Brazil and Mexico.

The study searches variables affecting the probability of reporting being happy in Colombia, Brazil and Mexico for the period 2010-2014. For this, a logistic specification ordered by country whose dependent variable is the reported happiness level expressed in categories was used. The conclusions of this empirical approach support that the level of income does not have a notable impact on reported happiness. In contrast,

variables such as marital status, health or the number of children have significant relevance on the probability of reporting being happy [9].

B. *Research 2: The Easterlin paradox in Spain*

Easterlin paradox refers to the fact that the growth of income per person (well-being in terms of goods and services) is not accompanied by a similar growth in the subjective feeling of satisfaction with life that the population declares in surveys for this purpose. In this study, data for Spain from 1980 to 2005 were analyzed to show that Easterlin Paradox is true. It was concluded that the Gross Domestic Product per person doubled during that range of years, but the average level of satisfaction of Spaniards with their life hardly increased [13].

III. METHOD

A. *Dataset*

The latest dataset of the World Happiness Report from Kaggle repository was consulted [12].

This dataset has the following structure shown in Table I.:

TABLE I
DATASET STRUCTURE

Country	Country name
Overall rank	Country ranking based on happiness score
Score	Individual personal happiness rating from 0 to 10.
GDP per capita	GDP per capita of each country in terms of purchasing power parity (PPP) (in USD)
Social support	Individual rating that determines whether, when you have problems, your family or friends would help you. Binary responses (0 or 1).
Healthy life expectancy	Healthy life expectancy at birth is based on data from the World Health Organization (WHO)
Freedom to make life choices	Individual rating that determines whether you are satisfied or dissatisfied with your freedom to choose what you do with your life. Binary responses (0 or 1).
Generosity	Generosity is the residual from the regression of the national mean of responses to the question "Have you donated money to a charity in the last month?" on GDP per capita.
Perceptions of corruption	Average of binary responses to two GWP questions: corruption in government and corruption in business.

B. *Extraction and use*

Google Colab environment and Python programming language were used for the extraction and statistical analysis of the variables involved. The libraries that were used in this environment were NumPy, Pandas, Matplotlib, and Seaborn.

C. *Exploratory data analysis*

As a first step to analyze and model the data, a statistical summary of the dataset was generated, which is shown in Table II. The target variable "Score" has a mean of 5,407, and the country with the highest score is Finland with 7,769 and the lowest is South Sudan with 2,853.

TABLE II
STATISTICAL SUMMARY OF ALL VARIABLES

	Overall rank	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	156.0000	156.0000	156.0000	156.0000	156.0000	156.0000	156.0000	156.0000
mean	78.5000	5.4071	0.9051	1.2088	0.7252	0.3926	0.1848	0.1106
std	45.1774	1.1131	0.3984	0.2992	0.2421	0.1433	0.0953	0.0945
min	1.0000	2.8530	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25%	39.7500	4.5445	0.6028	1.0558	0.5478	0.3080	0.1088	0.0470
50%	78.5000	5.3795	0.9600	1.2715	0.7890	0.4170	0.1775	0.0855
75%	117.2500	6.1845	1.2325	1.4525	0.8818	0.5073	0.2483	0.1413
max	156.0000	7.7690	1.6840	1.6240	1.1410	0.6310	0.5660	0.4530

Histograms were also generated to visualize data distribution of all variables in intervals, as shown in Fig. 1.

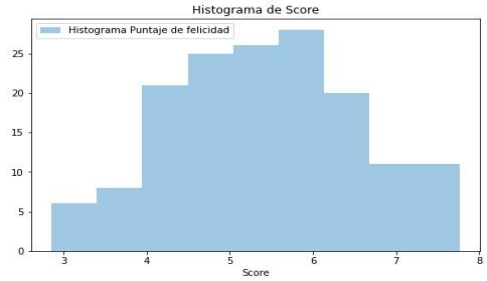

Fig. 1 Histogram of the distribution of the score variable

It was visualized that the Score variable has a standard normal distribution that is close to the Gaussian bell.

A normality test was also performed with the Shapiro-wilk test:

The following hypotheses were proposed:

Ho = Sample does not look Gaussian

Ha = Sample looks Gaussian

It results: *Statistics=0.965, p=0.001 then p<0.05*

Therefore, there is evidence to reject Ho which mentions that Sample does not look Gaussian.

Besides, GDP is evenly distributed worldwide, concentrated in the mean of its data, which is 0.9.

The variables Social support, Healthy life expectancy and Freedom to make life choices, have an asymmetric distribution skewed to the right, on the high values of each variable. This indicates that responses have a tendency to be positive for these variables, as shown in Fig. 2.

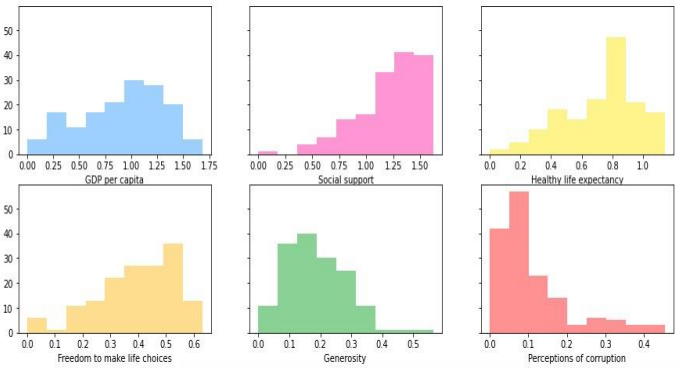

Fig. 2 Histogram of distribution of all variables

The dispersion of the data in each independent variable was analyzed by using box plots, as shown in Fig. 3. It can be seen that the variables Freedom to make life choices, Generosity and Perceptions of corruption have less dispersion, unlike the other variables that have greater dispersion, being GDP per capita the highest dispersion.

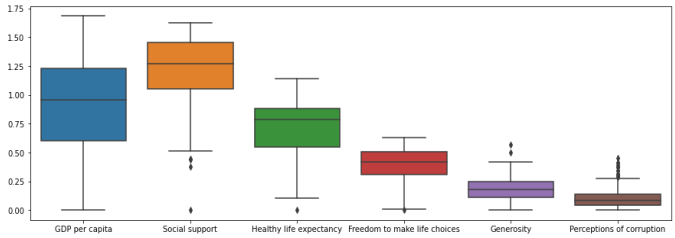

Fig. 3 Box plot of the independent variables

A visual way to analyze the relationship between two variables is through a linear regression graph. In this case, the dependent variable Score and the independent variables, as shown in Fig. 4.

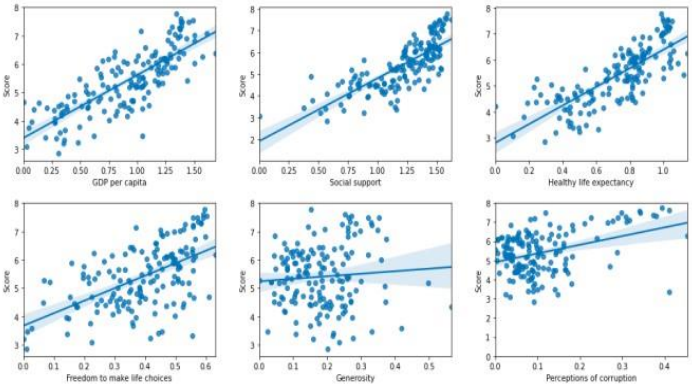

Fig. 4 Linear regression graph of the variables

The correlation was analyzed by using a heatmap, which shows the correlation between variables in a tabular way. See Fig. 5.

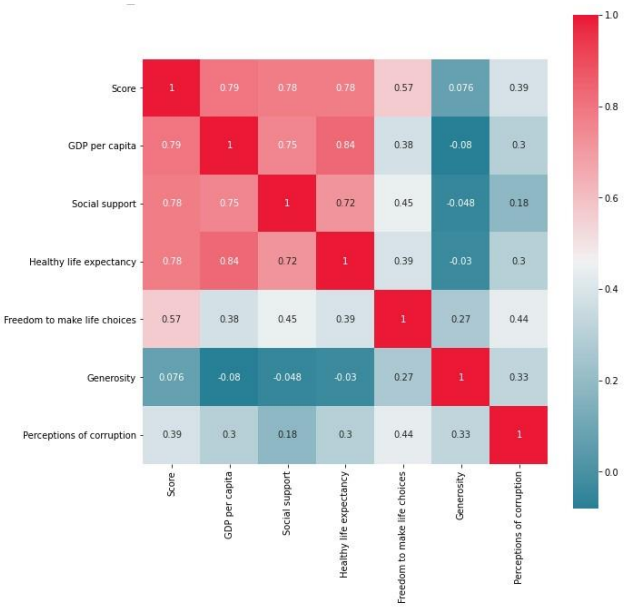

Fig. 5 Correlation heat map

It is observed that there is a high and direct correlation of the target variable Score with the variables GDP per capita, Social support and Healthy life expectancy, and the variable Generosity is a less important factor for happiness.

The correlation order is as follows:

1. GDP per capita (0.79)
2. Social support (0.78)
3. Healthy life expectancy (0.78)
4. Freedom to make life choices (0.57)
5. Perceptions of corruption (0.39)
6. Generosity (0.076)

IV. HYPOTHESIS

In the aforementioned researches, it is believed that income level does not necessarily influence people's happiness. Therefore, it is sought to corroborate that there are countries with high happiness but low GDP rate.

For this purpose, the mean μ of GDP was taken as a reference, which according to the previous analysis indicated to be 0.9 and it is also known that it has a normal distribution.

The following hypotheses were proposed:

Ho = All happy countries have a GDP rate>μ

Ha = There are happy countries that have a GDP rate<μ

A parametric Z-test was applied in which it was found that p_value <0.05.

z_score: 18.884

p_val: 1.54013e-79

Therefore, there is evidence to reject Ho which mentions that all happy countries have a GDP rate> μ.

To complement the analysis, a new three-level categorical variable Happiness_Category was generated, based on the variable called Score, under the following criteria:

- High Happiness: [6, 8>
- Medium Happiness: [4, 6>
- Low Happiness: [2, 4>

The dataset with the new variable Happiness_Category, is shown in Table III and Fig. 6 shows the Distribution of the new category by country.

TABLE III
DATASET WITH THE NEW TARGET VARIABLE

Rank	Country or region	Score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Score_Cat	Happines Category
1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393	7	1_High_Happiness
2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41	7	1_High_Happiness
3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341	7	1_High_Happiness
4	Iceland	7.494	1.38	1.624	1.026	0.591	0.354	0.118	7	1_High_Happiness
5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298	7	1_High_Happiness

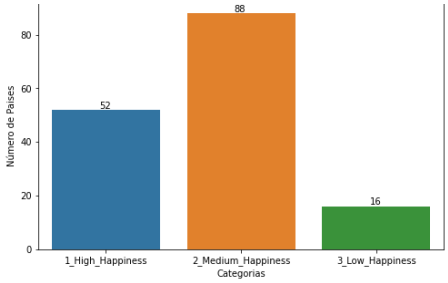

Fig.6.Distribution of the new category by country

A categorical scatter plot is shown below in Fig. 7, in order to have better visualization of the distribution of the new Happiness_Category with respect to the GDP variable. There are countries with High Happiness category and a GDP <μ mean of their values.

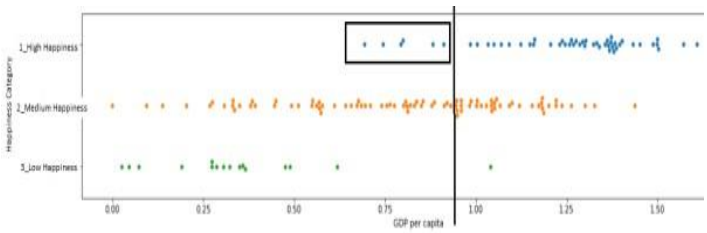

Fig. 7 Distribution of the new category by country according to GDP

As the cases of: Kosovo (0.882), Guatemala (0.800), El Salvador (0.794), Uzbekistan (0.745), Nicaragua (0.694). See Table IV.

TABLE IV
HAPPY COUNTRIES GDP <μ

27	Guatemala	6.438	0.800	1.269	0.746	0.535	1_High Happiness
35	El Salvador	6.253	0.794	1.242	0.789	0.430	1_High Happiness
41	Uzbekistan	6.174	0.745	1.529	0.756	0.631	1_High Happiness
43	Colombia	6.125	0.985	1.410	0.841	0.470	1_High Happiness
45	Nicaragua	6.105	0.694	1.325	0.835	0.435	1_High Happiness
46	Kosovo	6.100	0.882	1.232	0.758	0.489	1_High Happiness
50	Ecuador	6.028	0.912	1.312	0.868	0.498	1_High Happiness

This revalidates and rejects the hypothesis "All happy countries have a GDP rate> μ", from which it can be inferred that there are other factors that influence whether a country belongs to the High Happiness category. These would be Social support and Healthy life expectancy.

V. ANALYSIS OF RESULTS

After exploring the data, obtaining the characteristics of the dataset variables, and rejecting the null hypothesis, an analysis was performed by using Machine Learning algorithms.

To check the results provided by the models, r2-score or coefficient of determination were used, which allows to quantify the degree of adjustment between the measured data and the results of the model. It ranges between 0 and 1, when it acquires results closer to 1, the greater the model adjustment to the variable to be applied.

The other indicator was MSE (Mean Squared Error) which is a summary measure of the precision of the estimator.

Within this Machine Learning technique, regression, classification and group analysis algorithms were applied, for which the scikit-learn library was used, which is designed to work with these algorithms.

Supervised learning

A. Regression Algorithms

Their objective is to predict the Happiness Score through independent variables or factors.

Definition of target and predictor variables:

- Objective or target var.: Score
- Predictor var.: GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Perceptions of corruption, and Generosity.

Three types of regression algorithms were applied and the following r2-score values and errors were obtained:

a. Multiple Linear Regression:

Metric for regression:

mean absolute error:	0.499
mean squared error:	0.415
max error:	1.923
r2 score	**0.650**

b. XGBboots:

Metric for regression:

mean absolute error:	0.450
mean squared error:	0.394
max error:	1.658
r2 score	**0.667**

c. *Random Forest Regression:*

Metric for regression:

mean absolute error:	0.439
mean squared error:	0.333
max error:	1.380
r2 score	**0.719**

From the results, it was possible to identify which is the best prediction model for our objective variable Score, which was Random Forest algorithm with a better adjustment model, r2 = 0.719, as shown in Fig. 8.

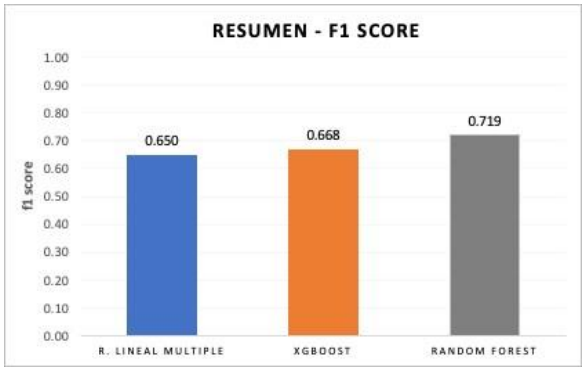

Fig. 8 Regression algorithm results

B. Classification Algorithms

Their objective is to predict the class or category of happiness through the independent variables or factors.

- Objective or target var.: a categorical variable "Happiness" was generated, under the following criteria:

  1 : [Score] > = 6

  2 : 4 <= [Score] < 6

  3 : [Score] < 4

- Predictor var.: GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Perceptions of corruption, and Generosity.

a. Gaussian Naive Bayes

	precision	recall	f1-score	support
class 1	0.889	0.667	0.762	12
class 2	0.667	0.824	0.737	17
class 3	0.000	0.000	0.000	3
accuracy			0.688	32
macro avg	0.519	0.497	0.500	32
weight avg	0.688	**0.688**	0.677	32

b. K Nearest-Neighbor

	precision	recall	f1-score	support
class 1	0.800	0.667	0.727	12
class 2	0.667	0.824	0.737	17
class 3	0.000	0.000	0.000	3
accuracy			0.688	32
macro avg	0.489	0.497	0.488	32
weight avg	0.654	**0.688**	0.664	32

c. Support Vector Machines

	precision	recall	f1-score	support
class 1	0.889	0.667	0.762	12
class 2	0.667	0.824	0.737	17
class 3	0.000	0.000	0.000	3
accuracy			0.688	32
macro avg	0.519	0.497	0.500	32
weight avg	0.688	**0.688**	0.677	32

It could be observed that the three algorithms have the same global prediction values 0.688, that is, they do not have much variability. On the other hand, they have a good prediction for class 1 and 2 but not for class 3.

Consequently, in order to decide which algorithm to take, the k-fold Cross Validation test was used, in which several training and testing tests (10 iterations) were performed.

It is observed in Table V that SVM algorithm has better prediction, with accuracy of 82% and variation of the tests of 0.099.

TABLE V
CROSS VALIDATION TEST RESULTS

cv = 10

		k-fold Cross Validation	
	f1-score	acc_mean	acc_std
GNB	0.6875	0.781	0.114
KNN	0.6875	0.814	0.091
SVM	0.6785	0.822	0.099

Below there is a graphical representation of the results of SVM algorithm. Fig. 9 shows the training set and Fig. 10 shows the test set.

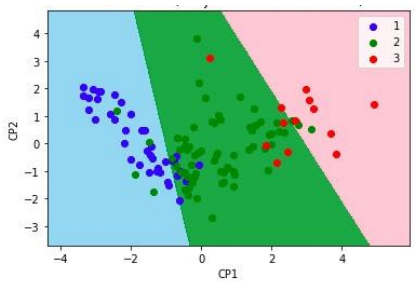

Fig.9.SVM training set

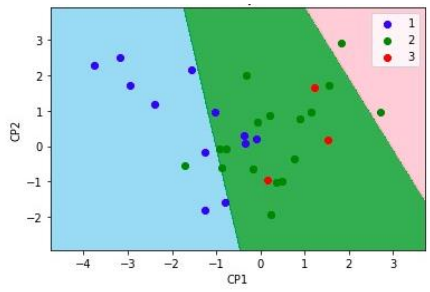

Fig.10 SVM test data

It is observed that it has good prediction for class 1 and 2, but not for class 3. One of the factors that may influence is the little amount of data available for that class.

Unsupervised learning

A. Clustering a. K-means

The elbow method was applied to find the cluster number and to be able to use it in the analysis with K-means, as shown in Fig. 11. This confirms the value used in classification analysis.

Cluster Number = 3

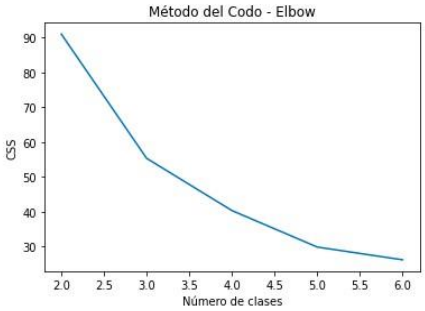

Fig.11 Cluster number calculation

In Fig. 12 and Fig. 13, it can be seen that GDP grows simultaneously with the variables Social support and Healthy life expectancy. Once again, the relationship of these three variables is checked.

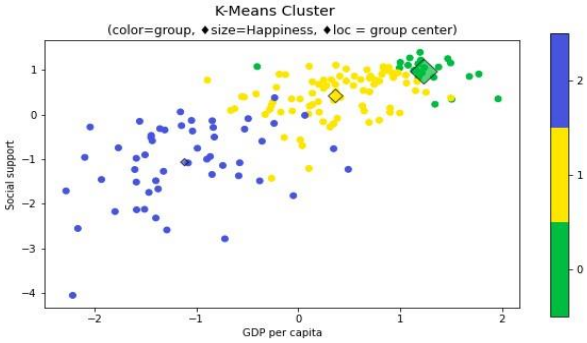

Fig.12 K-means GDP and Social Support

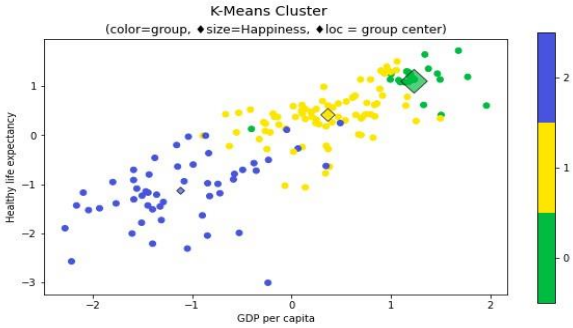

Fig.13 K-means GDP and Healthy life expectancy

Fig. 14 shows that the variable Freedom to make life choices is not decisive for GDP. There are countries that have high values for this variable (high freedom) and, in turn, low GDP.

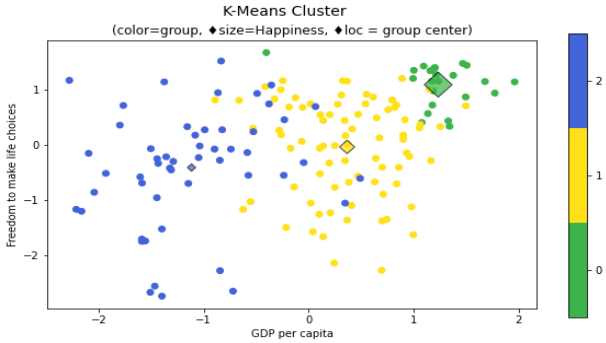

Fig.14 K-means GDP and Freedom to make life choices

Finally, Fig. 15 shows that the variable Perceptions of corruption is not decisive for GDP either. It should be emphasized that there are countries that have high values for this variable and also a high GDP.

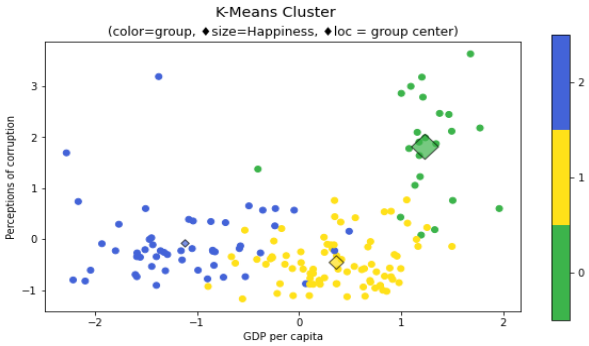

Fig.15 K-means GDP and Perceptions of corruption

B. Reduction of Dimensions a. Principal Component Analysis (PCA)

The World Happiness Report dataset shows six different variables to measure happiness score in 156 countries. PCA was applied to determine which components can represent happiness of the countries. Table VI shows that the first two

PC2 components represent approximately 74% of the global variance of the data, being the cut-off value for choosing the principal components of 70%. It is also clearly seen in Fig. 16.

TABLE VI
SUMMARY OF THE RESULTS OF THE EXPLAINED VARIANCE

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Explained_variance	0.498	0.238	0.102	0.093	0.044	0.026
Acum. Explained_variance	0.498	0.736	0.838	0.930	0.974	1.000

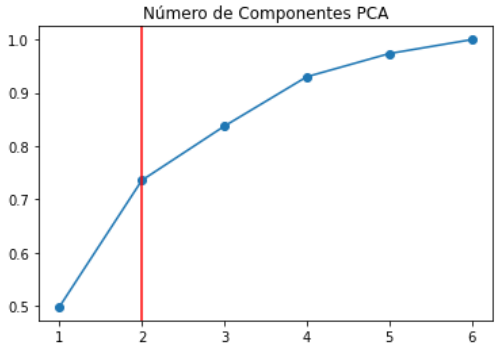

Fig.16 Number of PCA components

A Biplot is a global scatter plot aimed at representing both observations and variables of a multivariate data matrix on the same graph, helping to interpret the axes of the principal components while observing the location of individuals.

It is observed in Fig. 17 that coefficients of PC1 are positive for the six original variables; it has a positive score for the six variables. On the other hand, in the case of PC2, the countries with a high score in GDP, Social support, Healthy life expectancy are presented.

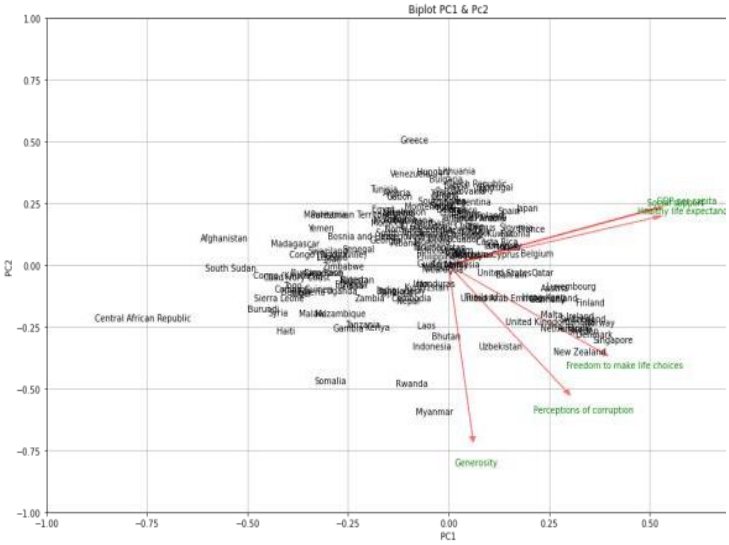

Fig.17 Global scatter plot (Biplot).

The biplot indicates that GDP per capita, Social support, Healthy life expectancy are highly correlated with each other. Likewise, Perceptions of corruption, Generosity, and Freedom to make life choices are among them.

Projecting a data point in the direction represented by an arrow gives the measures of those variables for that data value. For example, Myanmar has a higher value in Generosity than Finland, and the component chart confirms that Myanmar is closer to this vector than Finland.

VI. CONCLUSIONS

In this paper, an analysis of the factors that intervene in happiness was explained, based on data from the "World Happines Report".

The hypothesis proposed, which mentions that all happy countries have a GDP> μ, was rejected, when finding countries with a GDP lower than their arithmetic mean, as in the cases of Guatemala, El Salvador and Nicaragua.

It was also found that there are three latent factors or characteristics that influence happiness: GDP per capita, Social support, Healthy life expectancy.

By using Machine Learning, through the analysis of principal components, the dimensionality of the dataset could be reduced to 2 components with a representation of 73% of all the information and being able to reach 84% with 3 components.

Out of the three regression algorithms used to predict happiness score of a country, Random Forest proved to be more accurate, with 72%.

Similarly, by using clustering algorithms, it was possible to segment the countries into three groups with similar characteristics, with an accuracy of 82%.

REFERENCES

[1] Maslow, "A Theory of Human Motivation", Psychological Review, 50 (4), pp. 370-396, https://doi.org/10.1037/h0054346.
[2] United Nations. Happiness should have greater role in development policy, https://news.un.org/en/story/2011/07/382052
[3] E. Diener, E. Suh, R. Lucas and H. Smith, "Subjective well-being: Three decades of progress". Psychological Bulletin, 125(2), 276–302, 1999.
[4] E. Diener, M. Tamir and C. Scollon, "Happiness, life satisfaction, and fulfillment: The social psychology of subjective well-being.", 2006
[5] M. Seligman, "Florecer La nueva psicología positiva y la búsqueda del bienestar", March 2016.
[6] J. Ott, "Beyond Economics, happiness as a standard in our personal life and politics" pp71, 2020
[7] R. Ravina, J.Manchena and M. Montañés, "Happiness management en la época de la industria 4.0", Sep 2019.
[8] J.Helliwell, L Aknin, "Expanding the social science of happiness", Feb 2018,
[9] O. Poveda, "¿El ingreso influye en la felicidad de las poblaciones? Los casos de Colombia, Brasil y México", Jun 2019.
[10] The Harvard Gazette "Good genes are nice, but joy is better", April 2017
[11] World Happiness Report https://worldhappiness.report/

[12] Kaggle repository, dataset World Happiness Report 2019, https://www.kaggle.com/unsdsn/world-happiness?select=2019.csv
[13] J. Turmo, A. Rodriguez and O. Vara, “La paradoja de Easterlin en España”, April 2008.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE