Machine Learning & Statistical Applications to Countries' Standards of Living

Master of Quantitative Economics, University of California, Los Angeles

Youssef Ihab Mahmoud

Dr. Randall R. Rojas

05/26/2023

Table of Contents

2. Abstract

This project aims to investigate the utilization of machine learning and statistical techniques to examine the factors that affect the standards of living in countries, using the World Happiness Score as the determinant, dependent variable, of a country's standards of living. The factors used in this analysis are a mix of social, economic, geographic and political factors: the Freedom Index, Education Index, Perception of Corruption Index, Geographical Regions, GDP per Capita, Social Support Index, Health Life Expectancy, and the Human Development Index.

To gain insights and conclusions about the relationships among the variables, some statistical visualization techniques such as boxplots and a correlation heatmap were employed to convey the features' correlations' magnitude and direction with the Happiness Score.

Finally, Machine Learning Regularization and Ensemble models such as Lasso, Ridge, Elastic Net, Boosting and Random Forest were utilized to construct predictive models. The models used the factors to determine the extent of their importance on the Happiness Score, and whether they affect it in a positive or negative way. The performances of these models were evaluated using appropriate metrics such as root mean squared error and R-squared to assess their predictive accuracy.

After executing the techniques mentioned above, the analysis concluded that the most important factors, in order, were Health Life Expectancy, Social Support, GDP Per Capita, Human Development Index, Education Index, Freedom Index and Perceptions of Corruption. The regional factors showed little to no impact on the Happiness Score. All correlations with the Happiness score were positive except for Perceptions of Corruption and the Sub-Saharan African region. The best-performing Machine Learning model, Random Forest, achieved an R-squared of 83.3% and a root mean squared error of 4.03.

3. Introduction

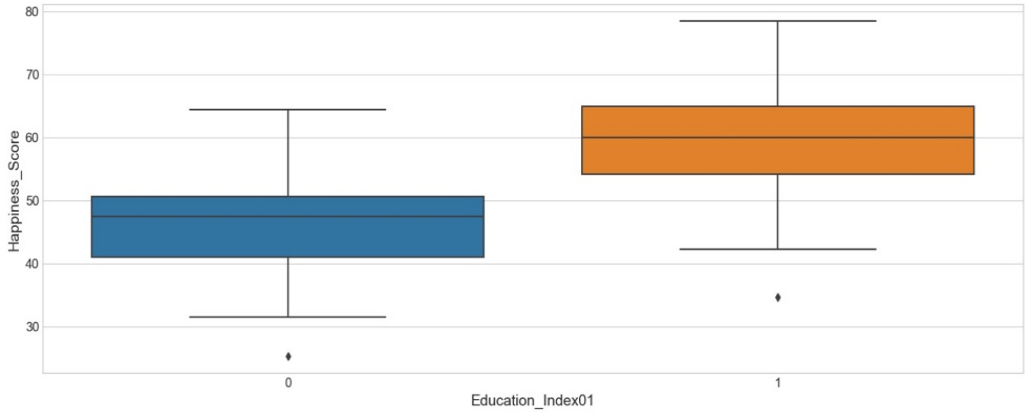
Understanding the factors that contribute to a country's standards of living is a complex and multidimensional task. Numerous socioeconomic, political, and environmental factors influence the well-being and happiness of individuals within a nation (Eichler,22). Studies on countries' standards of living usually focus on either Economic or Social aspects of life, however, there are a lot more factors that impact a country's standards of living. For this reason, this project will use statistical and machine learning techniques to show how various factors can play an impactful role in a country's standards of living.

Standards of living are typically higher in developed countries where the governments of these nations are economically powerful, provide good healthcare and education, allow more freedom, and are less corrupt (Boyle,2022). However, there is an important factor called the Human Development Index(HDI) which is a measure scored out of 100 that combines Health Life Expectancy, Education Quality, and Income Per Capita. The countries with the highest HDI were Norway (95.7), Ireland and Switzerland (95.5), Hong Kong and Iceland (94.9), and Germany (94.7) (Boyle, 2022). While the countries with lowest the HDI score were Niger (39.4), Central African Republic (39.7), Chad (3.98), Burundi and South Sudan (43.33), and Mali (43.4) (Eichler,2022). Four of the highest-ranked countries fall in Western Europe and all the lowest-ranked countries fall in the Sub-Saharan African region. This could indicate that there is also a regional factor that could play a role in a country's standards of living.

Economists usually use GDP per Capita as the measure of a country's standard of living, and it has been proven to have a strong correlation with a country's standard of living (Eichler,2022). However, this is a very narrow perspective that eliminates many other economic and non-economic aspects such as wealth inequality, and all the other factors mentioned earlier in the paper. This analysis will combine a variation of factors that could impact a county's standards of living which should prove that it is based on many different factors.

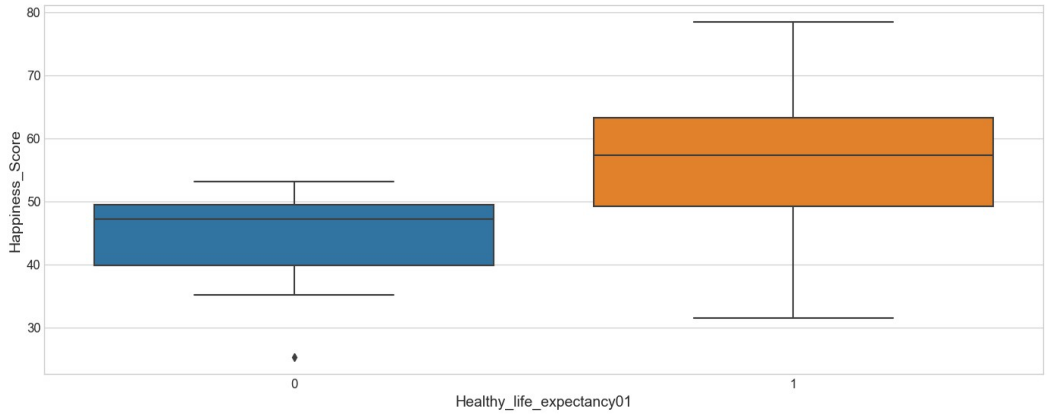
4. Definitions and Correlations

4.1 Education Index



This Boxplot is used to show the correlation between countries' Education Index score and their Happiness Score. The Education Index is a feature scored out of 100 to describe the quality of education in a country (UNDP, 2023). In the analysis, a binary variable for the Education Index was created to represent the difference in the Happiness Score between countries with an Education Index below 60 and countries with an Education Index above 60. The Boxplot indicates that countries with an Education Index score above 60 have a higher Happiness Score.

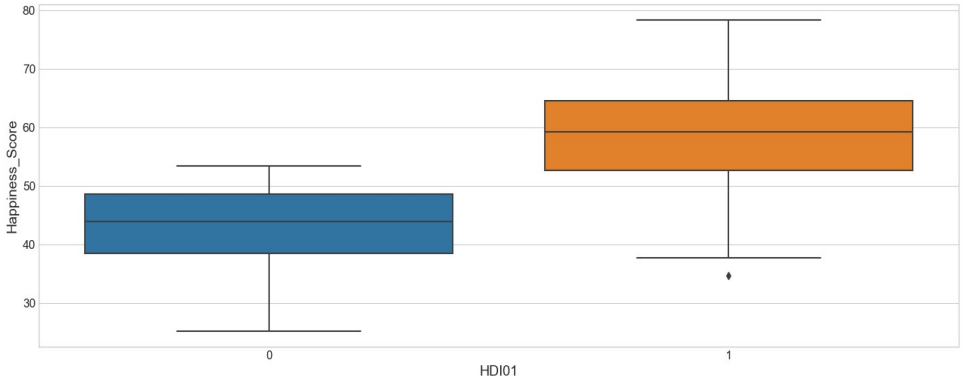
4.2 Health Life Expectancy



Health Life Expectancy is the average number of years a human is expected to be in a state of good health, accounting for years lived in non-optimal health conditions(WHO,2023). The Boxplot aims to present the correlation between countries' Health Life Expectancy and Happiness Score. In the

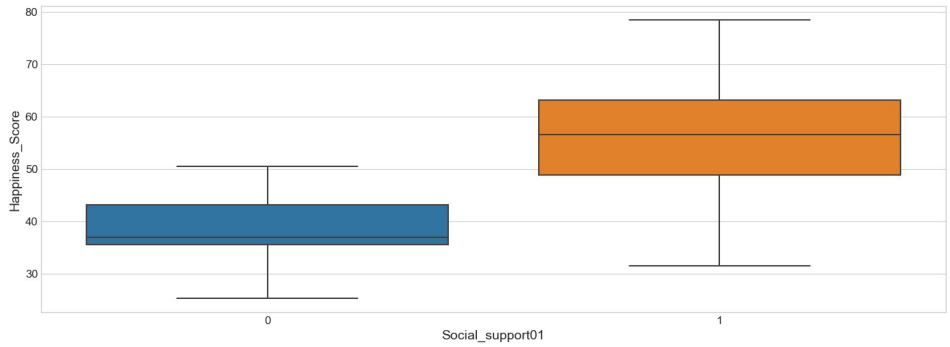
analysis, a binary variable for the Health Life Expectancy was created to represent the difference in the Happiness Score between countries with Health Life Expectancy below 55 and countries with Health Life Expectancy above 55. The Boxplot implies that countries with Health Life Expectancy above the age of 55 tend to have a higher Happiness Score compared to countries with Health Life Expectancy below 55.

4.3 Human Development Index



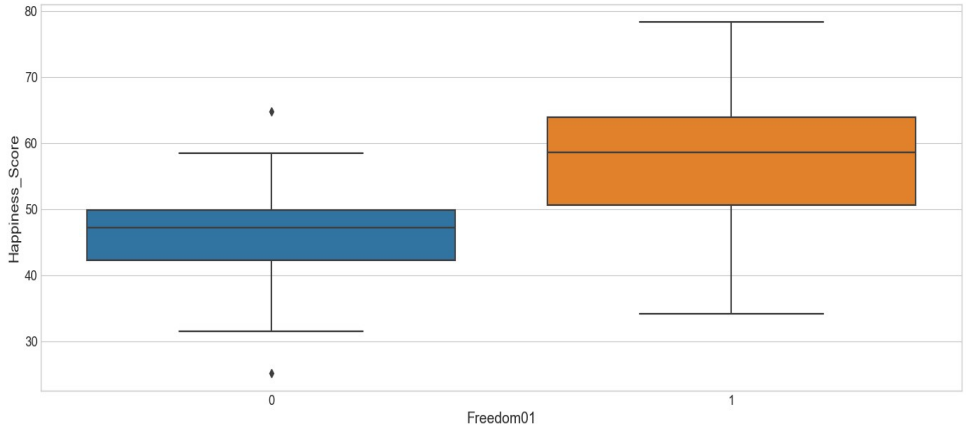
The Human Development Index (HDI) is a concise measurement used to evaluate important aspects of human development. It is calculated by taking the geometric mean of normalized indices of three essential dimensions of human development: having a lengthy and healthy life, being well-informed, and having a satisfactory standard of living (UNDP, 2023). In the analysis, a binary variable for the HDI was created to represent the difference in the Happiness Score between countries with HDI below 60 and countries with HDI above 60. The plot shows that countries with an HDI score above 60 have a higher Happiness Score than countries with an HDI score lower than 60.

4.4 Social Support



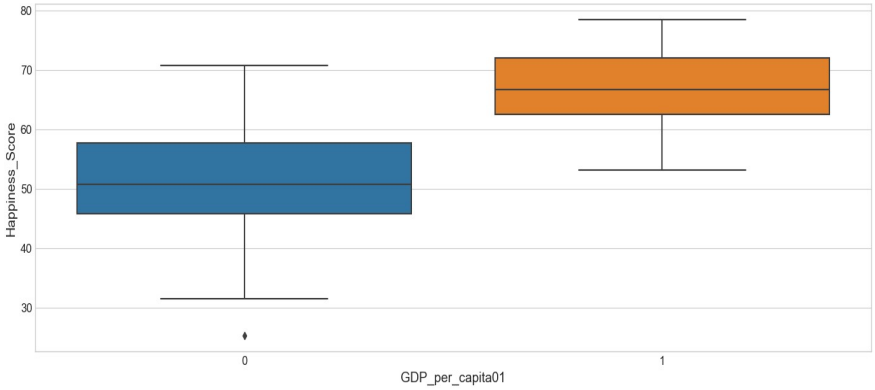
Social Support refers to having a network of people such as friends or family who can offer assistance and any kind of support during difficult life situations(University of Minnesota,2023). Social support has been psychologically and statistically proven to improve one's quality of life and overall well-being(University of Minnesota,2023). In the analysis, Social Support was scored out of 100 and a binary variable was constructed to show the discrepancies in the Happiness Score among countries with Social Support levels below 60 and countries with Social Support levels above 60. The plot conveys that countries with Social Support levels above 60 have a higher Happiness Score than countries with Social Support levels score lower than 60.

4.5 Freedom Index



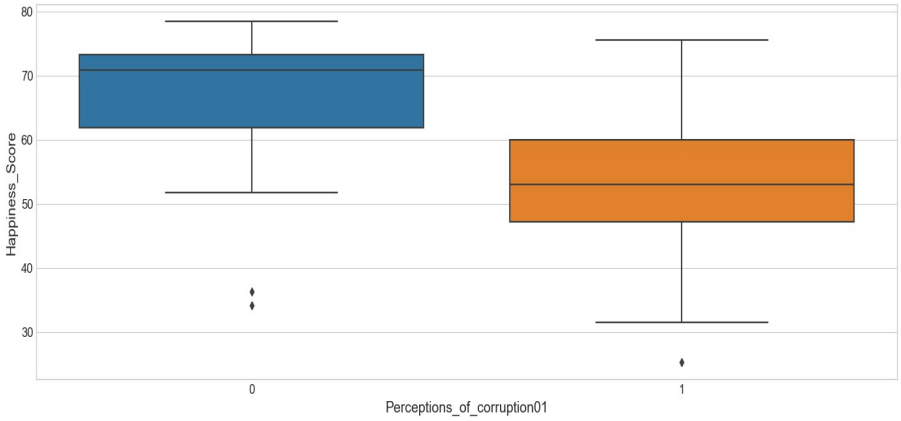
The Human Freedom Index explains the global human freedom in terms of personal, civil and economic freedom(Cato,2018). Freedom is scored out of 100 and a binary variable was created to separate countries with Freedom levels above and below 70. The plot shows that countries with a Freedom score above 70 tend to have a higher Happiness Score than countries with a Freedom score below 70.

4.6 GDP per Capita



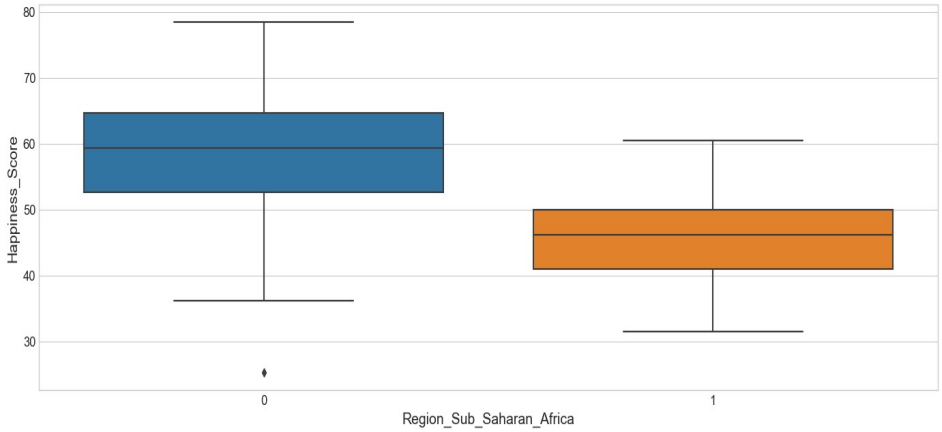
GDP Per Capita is the total value of goods and services annually produced in a country divided by the mid-year population of the country (Rathburn,2023). It is a global measure used to get a general understanding of a countries' Economic status. (Rathburn,2023) A binary variable was created to present the differences between countries with a GDP per Capita below and above $20,000. The plot indicates that countries with higher GDP per Capita tend to have a higher Happiness Score. The plot also shows that there are much more countries with a GDP per Capita below $20,000.

4.7 Perceptions of Corruption



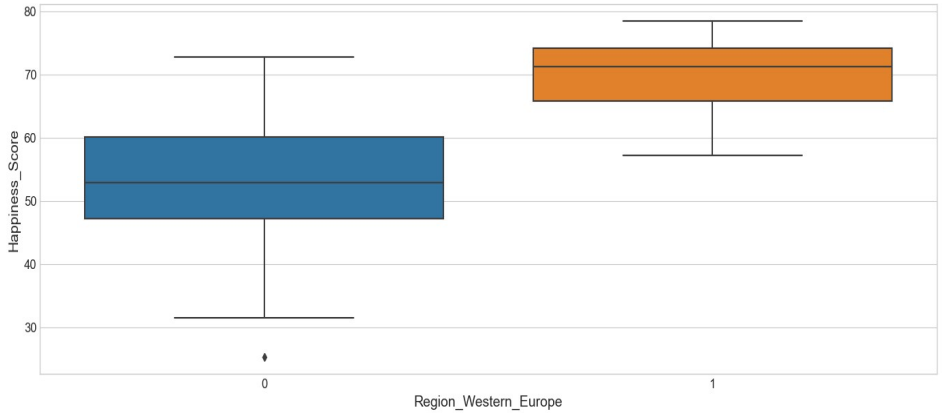
The Perception of Corruption Index is used to measure a government's level of corruption by country and is scored out of 100 (Kenton,2021). As high levels of corruption have been proven to be an obstacle to a country's political, social, and economic development, the plot supports this finding as it indicates that countries with corruption levels above 60 have a lower Happiness Score compared to countries with corruption levels below 60. (Kenton,2021)

4.8 Sub-Saharan Africa



This plot aims to show the differences in Happiness Scores between countries in Sub-Saharan Africa compared to all other regions. The plot indicates that countries in the Sub-Saharan Africa region have a lower Happiness Score compared to the rest of the World. This correlation aligns with a research study done by The Global Economy that showed that the average Happiness Score of an African country is 44.3, where most Sub-Saharan African countries were at the bottom of the list.(Eichler,2022).

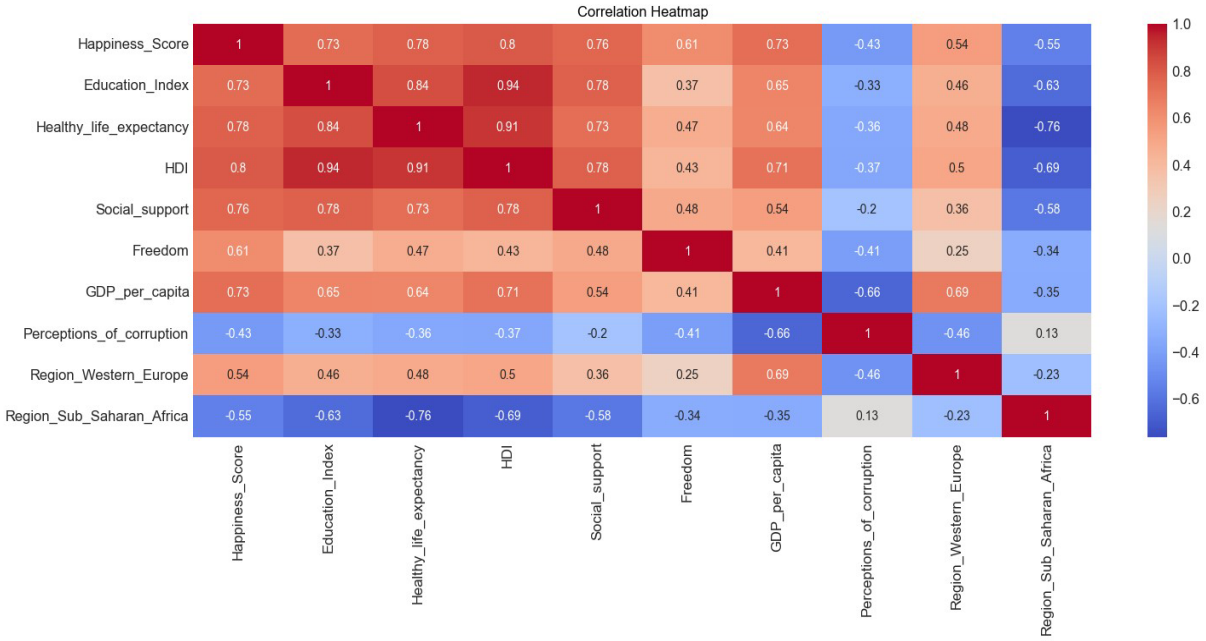
4.9 Western Europe



This plot explains the difference in Happiness Scores between Western European countries and the rest of the World. The Boxplot clearly infers that Western European countries have a much higher

Happiness Score compared to the rest of the World, with a median of approximately 72, without a single outlier among Western European countries.

4.10 Correlation Heatmap



The Correlation Heatmap is used to confirm the correlations, in terms of direction and magnitude, between all the features and the Happiness Score. The Red boxes indicate strong positive correlations between variables while the blue boxes indicate strong negative correlations. All the correlations in the Heatmap confirm the inferences made using the boxplots. An interesting takeaway from the Heatmap is that Western European countries have a positive relationship with all variables except Corruption, while countries in Sub-Saharan Africa have the exact opposite.

5. Regularization Methods

Regularization methods, also known as shrinkage methods, aim to fit the model including all the independent variables. The reason they're called shrinkage methods is because these models attempt to shrink the coefficients of the independent variables to zero, using penalizing parameters. Regularization methods are also to reduce variance and multicollinearity. Therefore, these methods are very useful for variable selection or in other words, showing variable importance, which is what this project aims to achieve. Before running any regularization method, the dataset should be split

into train and test sets, where the train set is used to predict the test set. In the analysis, the train size was set to 65% of the data. Another important tuning parameter that is adjusted before running a regularization method is the random state which is an integer that allows the execution of different train-test split each time the code is run. (James et al.,2021) In the analysis, the random state was set to 0 as it produced the most accurate results for the regularization models run.

5.1 Ridge Regression

As mentioned earlier, this regression aims to shrink the coefficient of the predictors to 0 using a shrinkage penalty usually referred to as lambda. As the value of lambda increases, the regression coefficient estimates will get closer to 0. In the analysis, a cross-validation was run to determine the optimal lambda for the model, 12.39. After running the regression using the best lambda, our model achieved a root mean squared error of 4.75 and a test R^2 of 76.59%. These metrics imply that on average our prediction for Happiness Score is off by 4.75 points and our predictors explain 76.59% of the variation in Happiness Score, implying that there is approximately 23% of the variation in Happiness Score is unexplained by the model.

Features	Education Index	Health Life Expectancy	HDI	Social Support	Freedom	GDP per Capita	Corruption Index	Western Europe	Sub-Saharan Africa
Coefficients	- 0.25	2.21	1.34	2.45	2.32	2.36	-0.88	0.56	-0.17

The Ridge Regression was not able to shrink any of the variables to 0 which is usually a disadvantage for Ridge Regression, it does not shrink variables to exactly 0 unless the lambda is approaching infinity (James et al.,2021). The direction of the coefficients, whether it is negative or positive, matches the directions found in the earlier correlation analysis used except for the Education Index which earlier showed a positive correlation with Happiness Score but now shows a negative correlation.

5.2 Lasso Regression

Lasso Regression also aims to shrink the predictors' coefficients to 0, however, it solves the problem with ridge regression, which is not shrinking variables to exactly 0. A Lasso regression produces simpler and more interpretable models that include the most important features only (James et al.,2021). Just like Ridge, Lasso also has a penalizing parameter that was cross-validated in the analysis and turned out to be 0.25 as the optimal lambda. After running the regression using the best lambda, our model achieved a root mean squared error of 4.85 and a test R^2 of 76.59%. These metrics imply that, on average, our prediction for Happiness Score is off by 4.85 points and our predictors explain 75.62% of the variation in Happiness Score, implying that there is approximately 24% of the variation in Happiness Score is unexplained by the model.

Features	Education Index	Health Life Expectancy	HDI	Social Support	Freedom	GDP per Capita	Corruption Index	Western Europe	Sub-Saharan Africa
Coefficients	0	3.082	0	2.79	2.24	3.25	-0.64	0	0

As expected, the Lasso Regression was able to shrink Education Index, Human Development Index (HDI), Western Europe, and Sub-Saharan Africa to 0 which could imply that these variables have no effect on the Happiness Score. In terms of RMSE and test R-squared, Ridge performed slightly better than Lasso, but Lasso was able to shrink four variables. Therefore, another regularization method will be used in the next step.

5.3 Elastic Net Regression

Elastic net regression utilizes both penalty parameters from Lasso and Ridge to shrink the unimportant features to zero and it improves the shortcomings of Lasso and Ridge(James et al.,2021). In the analysis, a cross-validation was run to determine the optimal penalizing parameter for the model, 0.22. After running the regression using the best penalizing parameter, the model achieved a root mean squared error of 4.73 and a test R^2 of 76.82%. These metrics imply that on average our prediction for Happiness Score is off by 4.73 points and our predictors explain 76.82% of the

variation in Happiness Score, implying that there is approximately 23% of the variation in Happiness Score is unexplained by the model. The elastic net performed slightly better than Lasso and Ridge as it produced a higher R-squared and a lower MSE.

Features	Education Index	Health Life Expectancy	HDI	Social Support	Freedom	GDP per Capita	Corruption Index	Western Europe	Sub-Saharan Africa
Coefficients	0	2.29	1.16	2.52	2.32	2.41	-0.82	0.46	- 0.04

Elastic net only shrank the Education Index to zero implying that it does not influence the Happiness Score. For the other features, the direction of the effect on the Happiness Score matches that of the correlation analysis performed earlier. Since all the regularization methods ran produced very similar accuracy results and different coefficient results, Ensemble models will be run to achieve results that are backed up by better scoring metrics.

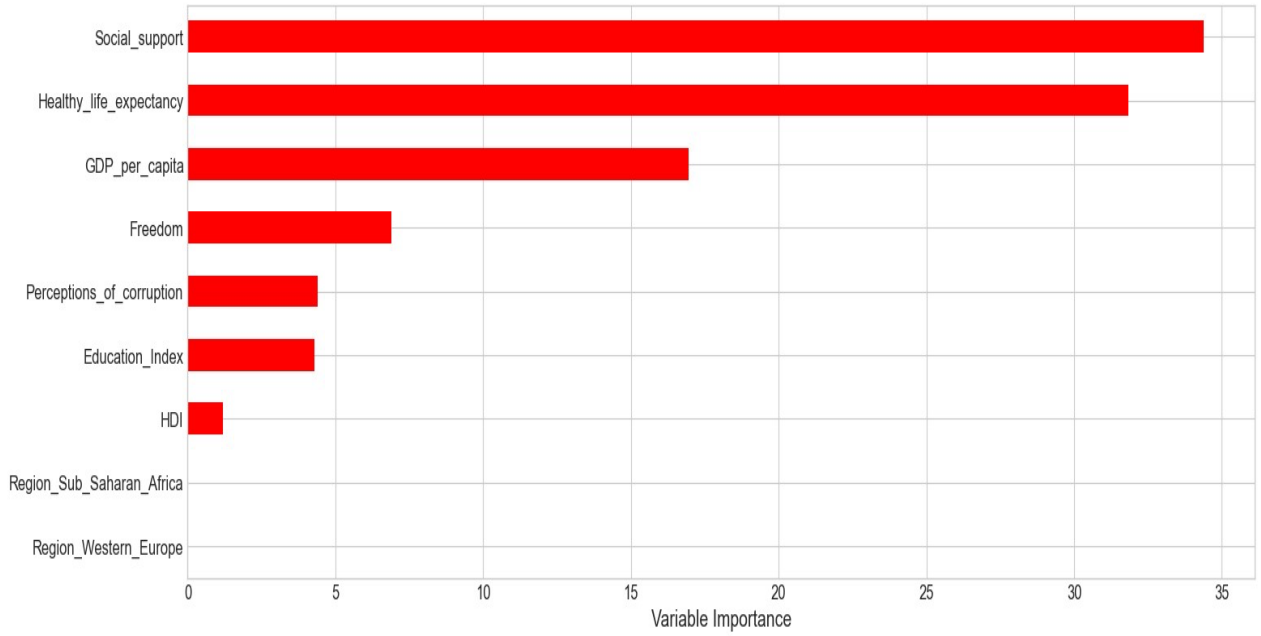
6. Ensemble Methods

Ensemble methods are machine learning methods that combine simple models with weak or inaccurate predictions, also known as weak learners, into a model that possesses strong predictive power. Ensemble methods can also be used to show which features have strong predictive power and which features have no impact on the target variable of interest. In the analysis, two ensemble methods will be used, Boosting and Random Forest (James et al., 2021).

6.1 Boosting

In the analysis, a cross-validation was run to determine the optimal penalizing parameter, also known as learning rate which is a small positive number between 0.01 and 1. The optimal learning rate was 0.03. After running the regression using the optimal parameters, the model achieved a root mean squared error of 4.07 and a test R-squared of 82.4%. These metrics imply that on average, our prediction for Happiness Score is off by 4.07 points and our predictors explain 82.4% of the variation in Happiness Score, implying that approximately 18% of the variation in the Happiness Score is

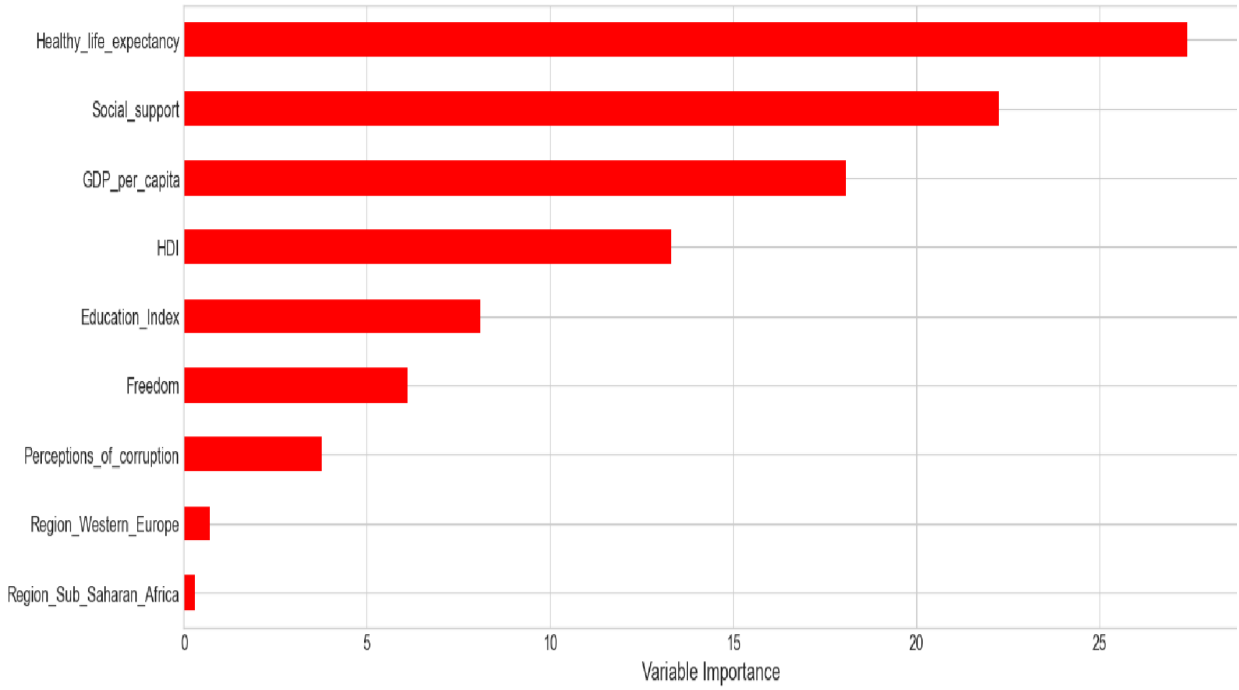
unexplained by the model. Boosting performed better than all regularization models, with a lower RMSE and a higher R-squared.



The figure above implies that all variables are important to a certain extent except for the region variables, which is consistent with the results from the Lasso. The Ridge and Elastic net regressions also reduced the region variables' coefficients below 1 which could also imply the insignificance of the variables. To confirm and finalize the results, one last ensemble method, Random Forest, will be discussed in the next section.

6.2 Random Forest

In the analysis, a cross-validation was run to determine the optimal maximum features to be included in the regression. After running the regression using the optimal tuning parameter, the model achieved a root mean squared error of 4.03 and a test R-squared of 83.3%. These metrics imply that on average, our prediction for Happiness Score is off by 4.03 points and our predictors explain 83.3% of the variation in Happiness Score, implying that approximately 17% of the variation in Happiness Score is unexplained by the model. Boosting performed better than all models, with a lower RMSE and a higher R-squared.



The figure above implies that most variables are important to a certain extent. The results are consistent with most of the other models. The region variables which are proven to be slightly important or unimportant in all models, are also proven to be slightly important in the best performing model, Random Forest. Most of the other variables show an acceptable level of importance that gives enough evidence to prove their impact on the Happiness Score.

7. Conclusion

The analysis concluded that in fact, countries' standards of living rely on a varied mixture of factors which could be strictly economic, strictly social, socioeconomic, and political. However, the analysis was not able to prove the strong importance between regional factors and a country's standards of living. Although the correlation analysis showed a significant correlation between regional variables and the Happiness Score, the machine learning models showed little to no impact on the Happiness Score. Not only the correlation analysis showed a strong pattern between regions and the Happiness Score, but also another empirical study concluded that Western European countries have the highest standards of living while Sub-Saharan African countries have the lowest standards of living (Eichler,2022). It is a takeaway that should be taken into consideration and used in further analysis to show why this correlation exists.

There were no discrepancies in the results of the correlation analysis, however, most of the machine learning models ran concluded that the most important factors were Health Life Expectancy, Social Support, GDP Per Capita, Human Development Index, Education Index, Freedom Index, and Perceptions of Corruption, with some discrepancies between the models' results. When there are minor discrepancies in results, it is better to rely on the model with the highest predictive accuracy and the lowest error, Random Forest. Examining the correlation analysis and the results of the Random Forest simultaneously, we conclude that all factors, varying in importance, have a positive correlation with the Happiness Score except for the Perceptions of Corruption and the Sub-Saharan African countries.

The discrepancies in the models ran and the unexplained variation in the Happiness score, which was 17%, implying that there are more factors to be considered to capture all or most of the factors that could impact a country's standards of living. Since Artificial Intelligence is a fast-paced progressing technology that could help fill in the gap in the analysis, more analyses about this topic should be performed to examine an overlooked, critical issue the world is facing today.

8. References

Cato Institute (2018, January 25). *The 2017 Human Freedom Index*. https://www.cato.org/multimedia/cato-daily-podcast/2017-human-freedom-index?gclid=Cj0KCQjwsIejBhDOARIsANYqkD2anHo2PH3VjYsSSMQOeP-cR2YbayeP02ezgSyD-CU5x7ftHwFLQs0aAhLFEALw_wcB

CIA (2021). *Unemployment Rate*. https://www.cia.gov/the-world-factbook/field/unemployment-rate/country-comparison

Eichler, R., & Boyle, M. (2022). *Standard of Living Definition, How to Measure, Example*. Investopedia. https://www.investopedia.com/terms/s/standard-of-living.asp

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Kenton, W. (2021, July 30). *Corruption Perceptions Index (CPI): Definition, Country Rankings*. Investopedia. https://www.investopedia.com/terms/c/corruption-perception-index.asp

Rathburn, P. (2023, April 23). *GDP Per Capita Defined: Applications and Highest Per Country*. Investopedia. https://www.investopedia.com/terms/p/per-capita-gdp.asp

SCHMITT, K. (2022, September 14). *Https://www.Investopedia.Com/terms/h/human-development-index-hdi.Asp*. Investopedia. https://www.investopedia.com/terms/h/human-development-index-hdi.asp

The World Bank (2021). *GDP Per Capita*. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=AF

Transparency International (2021). *CORRUPTION PERCEPTIONS INDEX*. https://www.transparency.org/en/cpi/2021

United Nations (2023). *Human Development Index (HDI)*. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

University of Minnesota (2023). *Social Support*. Taking Charge of Your Health and Well-Being. https://www.takingcharge.csh.umn.edu/social-support#:~:text=Social%20support%20means%20having%20friends,buffer%20against%20adverse%20life%20events.

World Data (2021). *Average Income Around the World*. https://www.worlddata.info/average-income.php

World Health Organization.(2023). *Healthy life expectancy (HALE) at birth*. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/66