Can an Algorithm Help With Your Happiness?

Final report

IS 733 Data Mining
Fall 2020

Submitted by

Antonios Xenakis, Leena Singh, Nischay Gupta, Daniel Adjei

- Abstract

Happiness, the most important aspect that will lead us to a successful and peaceful life. But, Is there any way to measure it, or find what are the reasons behind people's happiness? To measure the magnitude of happiness, several research studies have also been conducted. We worked on various machine learning algorithms and models like Linear Regression, Ridge Regression, Decision Tree, Support Vector Regression, Random Forest etc to find more of these parameters and their contribution to Happiness. Our work is about calculating the happiness score based on different parameters and depending on these, most and least happy countries have been predicted. We will conclude with the most effective parameter and the algorithm with best results for desired output.

- Introduction

The purpose of this project is to access the overall mood of a nation as well as give a glimpse into how it is evolving over time using an algorithm. Even though it may seem difficult or vague to gauge the happiness of a nation, this research uses tangible metrics to predict happiness.

The motivation for choosing this project work is to discover which important factors will lead us to a happy life. Various Machine Learning algorithms, to calculate and predict the happiness score were used, and there was comparison as to which is the best among them. Hence, people around the world can use the happiness score and focus on these parameters in order to achieve happiness.

Metrics used include the GDP per capita which is the measure of a country's economic output that accounts for its number of people. Another metric considered is the social support of friends and family to turn to in times of need or crisis to give you a broader focus and positive self image**.** Health life expectancy which is the average number of years that a newborn can expect to live in full health or not hampered by disabling illness or injuries is one of the metrics being looked at to measure the happiness score of a country. Freedom of an individual to make life choices from at least two options unconstrained by external factors and the quality of being kind and generous are two factors that are being considered in this project. Finally, perceptions of corruption using the Corruption Perception Index (CPI) which is annually published by the Transparency international which ranks countries by their perceived levels of public sector corruption is also considered in measuring the happiness score of a country.

The paper is divided into various sections. The first section being abstract which contains the summary of the paper. Followed by an introduction to the paper. Related work that clearly describes a previous work surrounding the happiness score is also included in the paper. The next section contains information concerning the dataset which was used for the project followed by the pre-processing activities performed on the data in order to generate clean and relevant data for processing and easy interpretation by the algorithm. A section which includes exploratory data analysis, methods such as LinearRegression, Decision Tree, Support Vector Regression, Bayesian Ridge, Ridge, Gradient Boosting and Random forest. Also explained are the approaches used to achieve the goal of the project, along with its results. The conclusion based on the outcome of the results. The future scope of this paper is summarized and the last section contains out references used for this paper.

- Related Work-Literature Review

According to Cummins (Cummins et al. 2003), there are multiple instruments for calculating satisfaction with life. To settle on one of them depends on how their psychometric properties fit the aim of the study.

Researchers have different views on how to measure happiness, whereas in this paper the metrics used are both subjective and objective, others believe happiness is solely subjective. For instance, a group of researchers believe that estimating situational factors such as becoming a spouse or a parent should be the measurement of one’s happiness (Kohler et al.2005). Other groups of researchers believe that genetics is a single major factor for achieving high happiness index, this conclusion was arrived by a study involving several thousands of twins which indicated that genetic influences are the single major factor explaining 44–52 % of the variance in subjective well-being (Lykken and Tellegen1996). That same study, (Lykken and Tellegen 1996), proved that circumstances accounted for a surprisingly small amount of variance, for example, less than 2 % for educational attainment, socio-economic status, or income, and less than 1 % for marital status.

Multiple metrics are used in this project because historically, governments have used GDP as a main indicator of happiness (Ovaska and Takashima, 2006). This is an incomplete measurement and does not tell the full story because there is a disconnect between GDP and personal income and it is likely for a GDP of a nation to rise while per capita income declines for the vast majority of the population income declines (Layard, 2005). Other researchers posit that happiness indicators, just like the factors considered in this paper, include personal relationships, economic freedom, health, education and income distribution (Ovaska and Takashima, 2006).

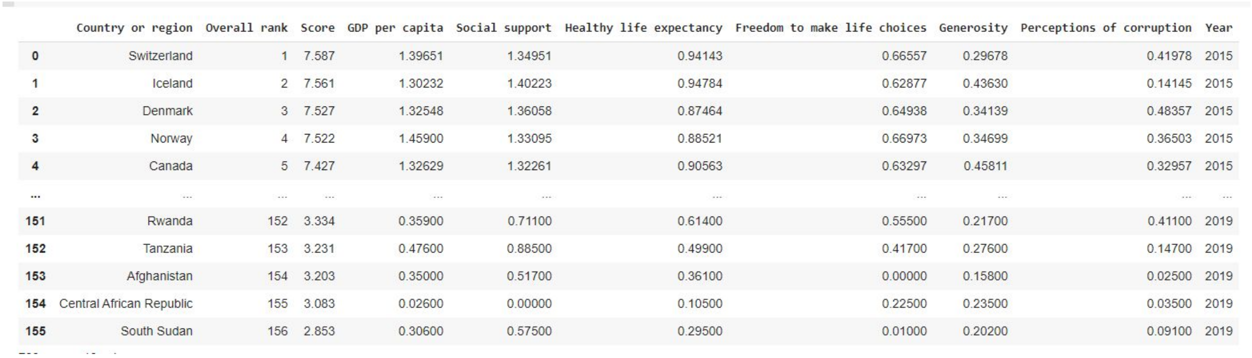
- Dataset

The data is from the World Happiness Report which is a survey of the state of global happiness. The data consists of information for the years 2015-2019. Each year has some different attributes besides the country and the overall rank which are consistent over the five years. The dataset is downloaded from kaggle at https://www.kaggle.com/unsdsn/world-happiness in a comma separated variable (CSV) format with a total size of 37 kb. The total rows for the years under review i.e. (2015 - 2019) come up to over 700 rows.

The dataset contains over 158 countries from Sub-Saharan Africa, Central and Eastern Europe, Latin America and Caribbean, Western Europe ,Middle East and Northern Africa regions. Common attributes include Happiness Rank, which gives the rank of the country based on the Happiness Score. Happiness Score metric which measured a specific year by asking the sampled people the question: “How would you rate your happiness on a scale of 0 to 10 where 10 is the highest.” The standard error of the happiness score. The Economy which shows the extent to which GDP contributes to the calculation of the Happiness Score. Family; the extent to which Family contributes to the calculation of the Happiness Score. Health; the extent to which Life expectancy contributed to the calculation of the Happiness Score. Freedom; the extent to which Freedom contributed to the calculation of the Happiness Score and finally Trust; the extent to which Perception of Corruption contributes to Happiness Score.

- Pre-processing

Steps used in pre-processing include merging all the five year dataset and normalising the column names in order to have uniformity and finding missing values and handling them by replacing the null values by dropping them if they may not cause a significant impact on the results.



	Country or region	Overall rank	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Year
0	Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	0.29678	0.41978	2015
1	Iceland	2	7.561	1.30232	1.40223	0.94784	0.62877	0.43630	0.14145	2015
2	Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	0.34139	0.48357	2015
3	Norway	4	7.522	1.45900	1.33095	0.88521	0.66973	0.34699	0.36503	2015
4	Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	0.45811	0.32957	2015
...	...	...	...	...	...	...	...	...	...	...
151	Rwanda	152	3.334	0.35900	0.71100	0.61400	0.55500	0.21700	0.41100	2019
152	Tanzania	153	3.231	0.47600	0.88500	0.49900	0.41700	0.27600	0.14700	2019
153	Afghanistan	154	3.203	0.35000	0.51700	0.36100	0.00000	0.15800	0.02500	2019
154	Central African Republic	155	3.083	0.02600	0.00000	0.10500	0.22500	0.23500	0.03500	2019
155	South Sudan	156	2.853	0.30600	0.57500	0.29500	0.01000	0.20200	0.09100	2019

Figure 1: Final Table Sample

Unique attributes that are not included in every year identified and analysed to understand how their removal would impact the accuracy. To identify columns that are not needed and remove them we created a correlation matrix individually for every year at first to see the significance of each column in relation to the rest and then we generated a new correlation matrix based on the final table as seen in Figure 2. We dropped the columns Country of Region, Overall Rank and Year before we move on to train our models.

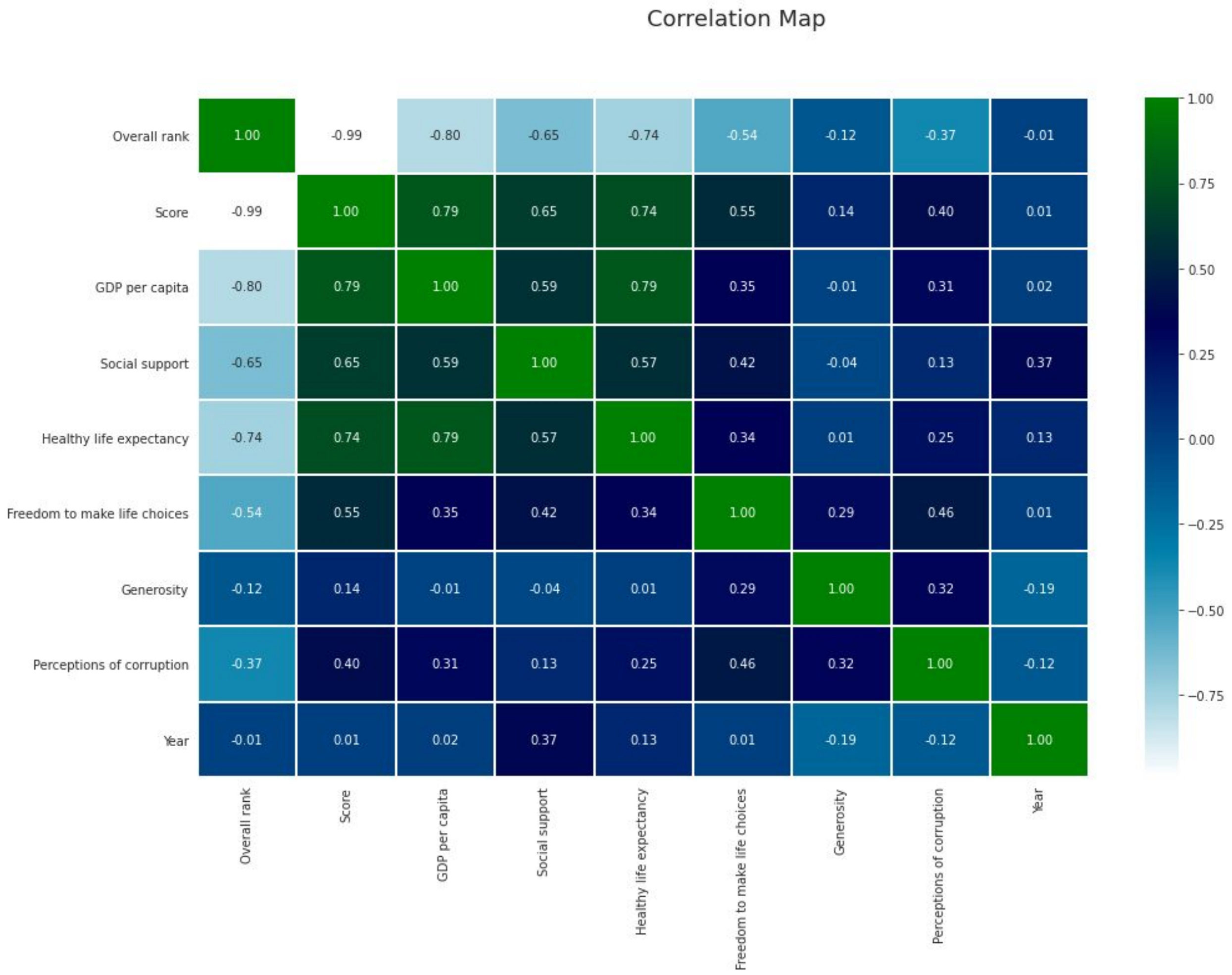


Figure 2: Correlation Map

- Exploratory Data Analysis
  After pre-processing and identifying the correlation between features we moved to data analysis to gain some insights. As a first step we tried to see the difference in years of different factors in comparison to the average score of each year. As we can see in Figure 3 throughout all the years the factors that seem to have the highest numbers are money and family.

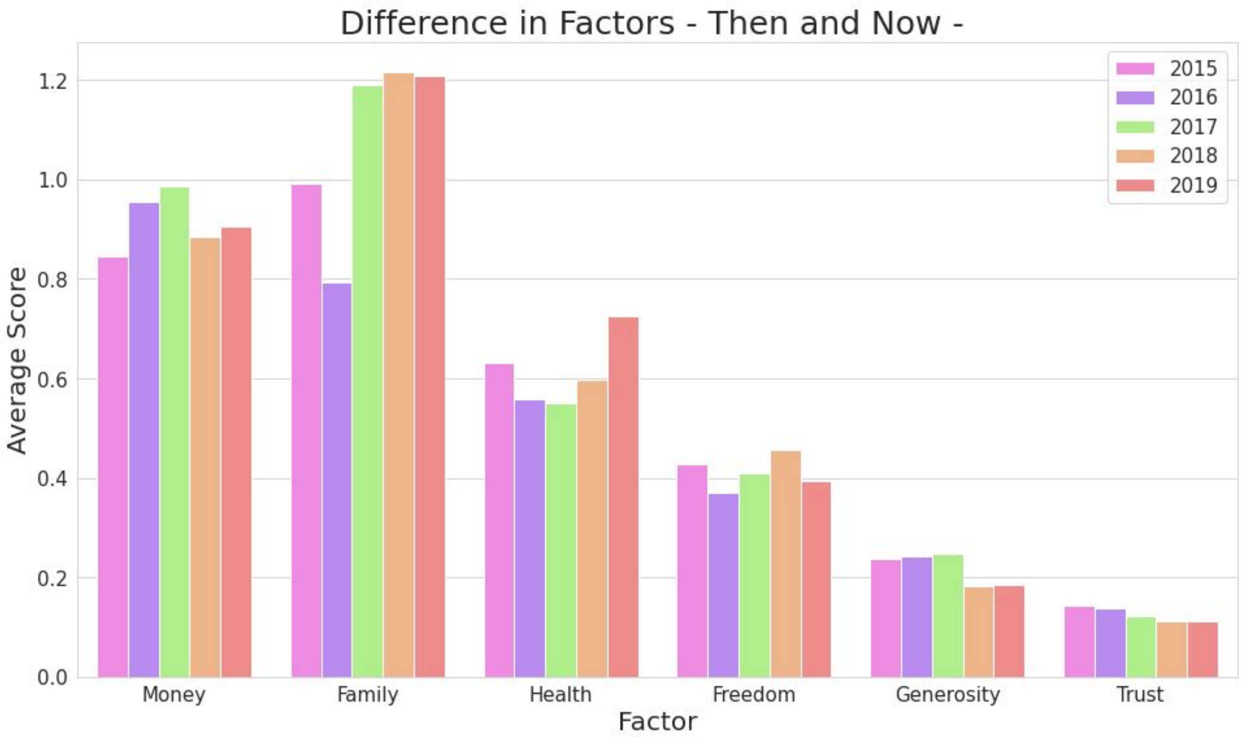


Figure 3: Difference in Factors from All Years

Since, we identified some of the most important factors then we found which countries had the highest happiness score in the most recent year. As seen in Figure 4, Finland, Denmark and Norway are the top three leading countries in 2019. Most of the countries with the highest score being in Europe besides New Zealand and Canada.

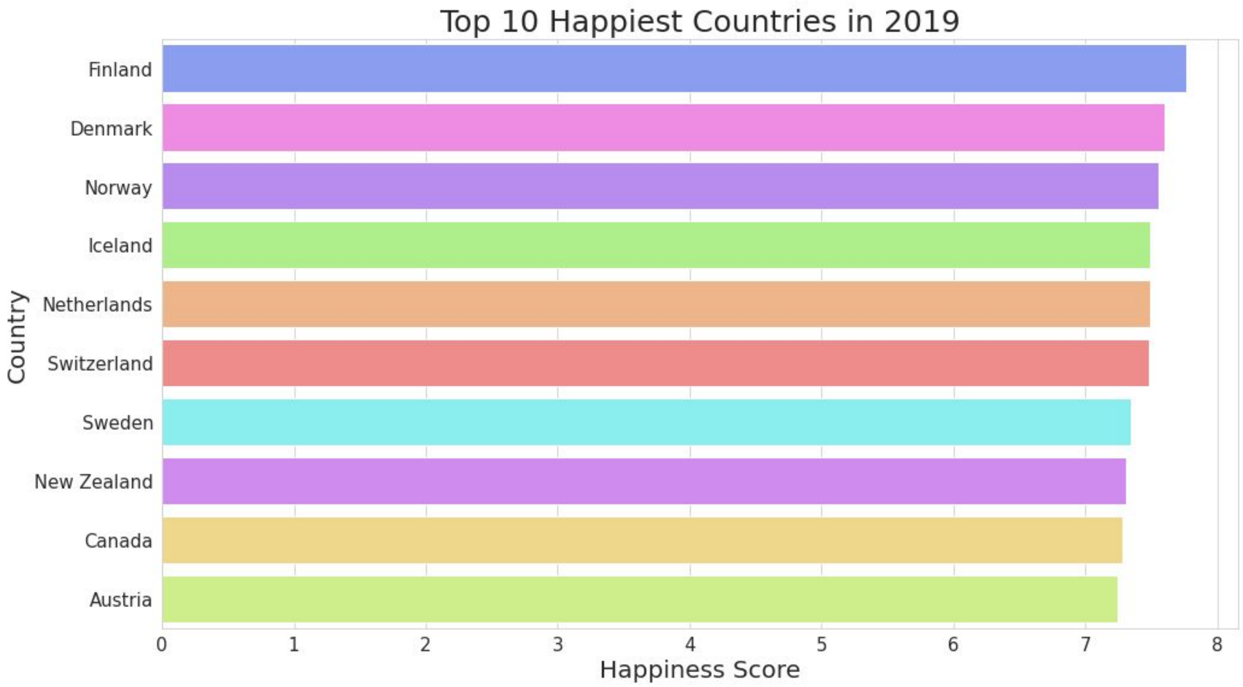


Figure 4: Top 10 Happiest Countries in 2019

After reviewing the countries with the highest scores and the most important factors we wanted to get a better understanding of the distribution of the average happiness score. The country score average was calculated against its density and we can see peeking around the ~4.5, ~5.3 and ~6.3. With that observation we can say that the most seen scores in our data-set are moving in that range and scores with a range ~7.5 coming next and ~3.5 having the lowest density.

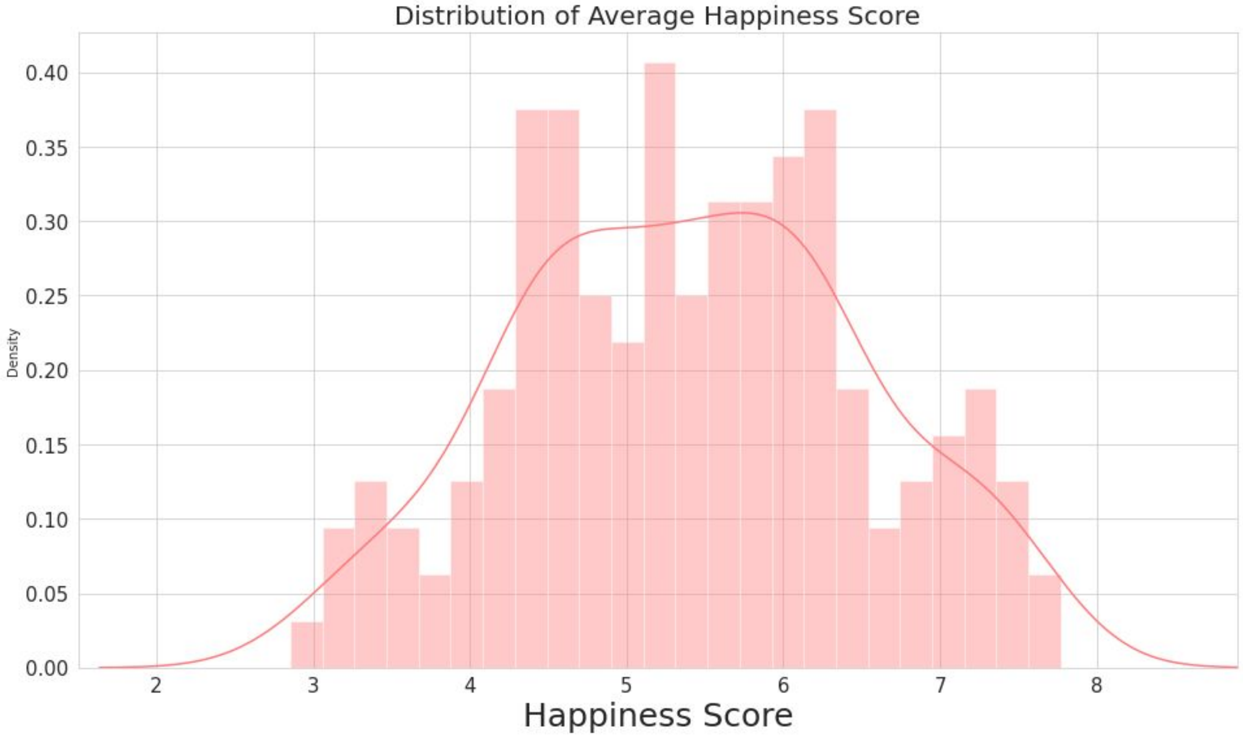


Figure 5: Average Happiness Score Distribution

- Methods

  ▪ Linear Regression

  An approach to model a relationship between dependent and independent variables. It is a statistical approach which is done for the purpose of predictive analysis. It does not make predictions for the categorical variables but continuous numerical variables. Depending on the number of input variables, it has two methods: Simple and Multi linear regression.

  ▪ Decision Tree

  Decision tree is an approach that the data splits based on certain parameters in two parts, nodes and leaves. Leaves are where a decision is being made or a final outcome and nodes are where the data are getting split.

  ▪ Support Vector Regression

  Support Vector Regression (SVR), is a supervised machine learning technique that uses a binary classification algorithm or used for regression analysis when a prediction for numerical data is needed.

▪ Bayesian Ridge

Bayesian Ridge Regression is a mechanism which can take the undistributed/poor data using probability distribution instead of point estimates i.e. the response is not considered as the single best value but estimated from probability. Here, the value is provided through spherical Gaussian..

▪ Ridge

Ridge regression is the method used for analyzing the data related to multiple regression having multicollinearity. It performs better when we talk about long term predictions. It is similar to linear regression but in the model we put little bias which results in remarkable fall in variance. By adding the bias, it minimized the standard error.

▪ Gradient Boosting

Gradient boosting is a machine learning technique for regression in a form of an ensemble of weak prediction models, typically decision trees. Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function to current residuals by least squares at each iteration (Jerome H. Friedman, 2002).

▪ Random Forest

Random Forest, builds multiple decision trees and then merges them together to build a prediction model. A similar approach to decision trees with the difference that random forest consists of multiple trees of random sampling.

- Approach

  Our data-set is ready to start training the models, first we assigned the features needed to train the model and excluded our label from the axis. Next step was to split our data into train and test sets. Keeping the test-set size to 20% and the rest 80% the train-set .

  In our approach we used seven different models, Linear regression, Decision Tree, Support Vector Regression, Bayesian Ridge, Ridge, Gradient Boosting and Random Forest and tried to identify which one would have the most successful predictions. All of our approaches are supervised machine learning models and in our case we will use regression since our target variable consists of numerical values.

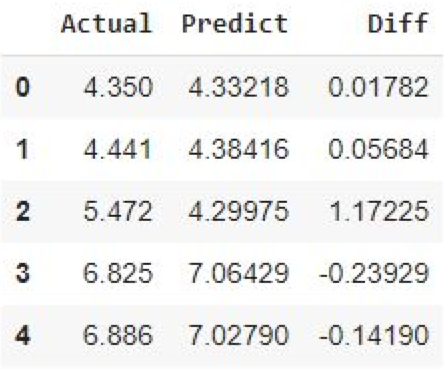
  Initially we applied a linear regression approach, to model a relationship between our attributes and establish a baseline for our results by trying to predict the score, achieving a Root Mean Square Error (RMSE) of 56.14%. Then we moved on to implementing a Support Vector Regressor (SVR) model and Random Forest. SVR achieved a RMSE of 57.83% and Random Random Forest a 50.96%, which made it our most accurate prediction amongst these three algorithms. After experimenting with these three models we also implemented Ridge Regression, Decision Tree Regression, Bayesian Ridge and Gradient Boosting Regressor, achieving RMSE of 56.15%, 65.65%, 56.16%, 50.75% respectively. Similar results to the previous models, having Gradient Boosting Regressor being the best performing approach.

- Results
  Evaluating our results we used different metrics such as, Mean Absolute Error, Mean Square Error, Root Mean Square Error, Coefficient of Determination (R^2) and Variance Score. As we see in Table 1 comparing the performance of SVR and Random Forest Regression, Random Forest is performing slightly better with a difference around ~6%.

	Support Vector Regression	Random Forest Regression
Mean absolute Error (MAE)	0.4365	0.3975
Mean Square Error (MSE)	0.3344	0.2597
Root Mean Square Error (RMSE)	0.5784	0.5096
Variance Score	0.73	0.79

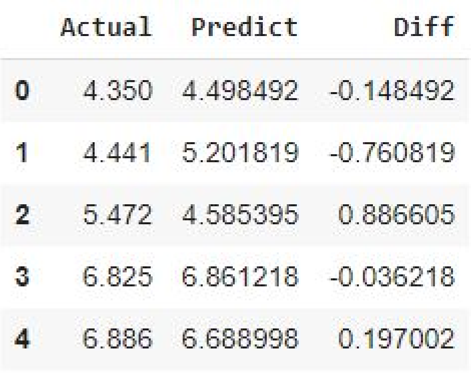
Table 1: SVR and RFR Results

In Table 2 and Table 3 underneath we have an example of the actual values, predicted values and the difference between them.



	Actual	Predict	Diff
0	4.350	4.33218	0.01782
1	4.441	4.38416	0.05684
2	5.472	4.29975	1.17225
3	6.825	7.06429	-0.23929
4	6.886	7.02790	-0.14190

Table 2: Actual and Predicted Values
for RF



	Actual	Predict	Diff
0	4.350	4.498492	-0.148492
1	4.441	5.201819	-0.760819
2	5.472	4.585395	0.886605
3	6.825	6.861218	-0.036218
4	6.886	6.688998	0.197002

Table 3: Actual and Predicted Values
for SVR

In the following graphs, Figure 5 and Figure 6, we constructed a graph with predicted and actual values, closer to the line the more accurate the prediction. Comparing the two graphs we can see Random Forest is more dense closer to the line.

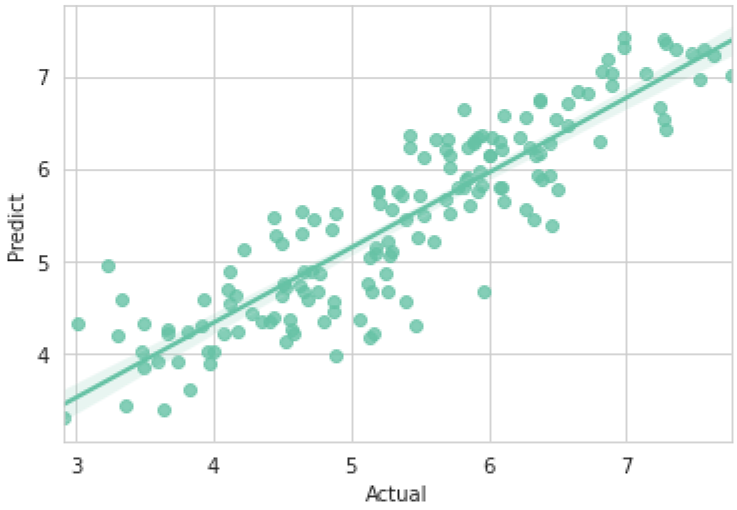


Figure 6: Random Forest Graph

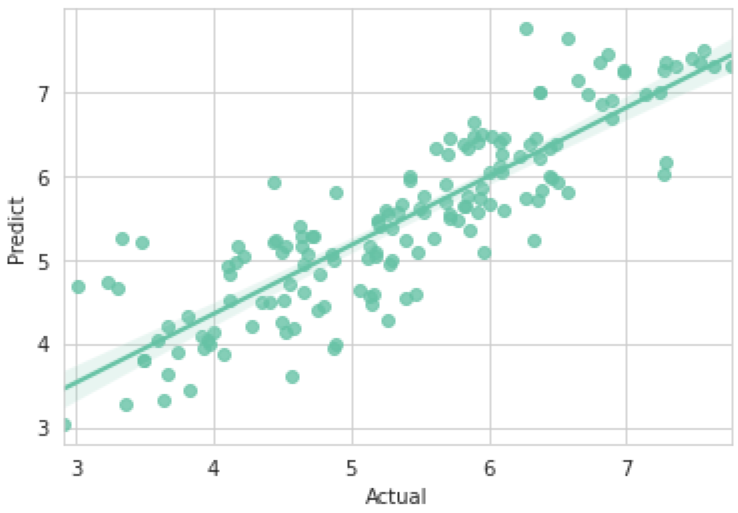


Figure 7: Support Vector Regressor Graph

In the following table we have evaluations of the rest of the models, comparing them on root mean square error and coefficient of determination. From the RMSE, we can tell that Gradient Boosting Regression outperformed all the models with 50.74% achieving ~16% difference from the worst performed model, Decision Tree.

	Linear Regression	Ridge Regression	Decision Tree	Bayesian Ridge Regression	Gradient Boosting Regression
Root Mean Square Error (RMSE)	0.5615	0.5615	0.6565	0.5616	0.5074
Coefficient of Determination (R^2)	0.7465	0.7465	0.6535	0.7464	0.7931

Table 4: Results

In the Figure 7 given below, a comparison is done using R-square score between several machine learning algorithms. The Gradient Boosting proving itself the most and the least preferred method will be the Decision Tree.

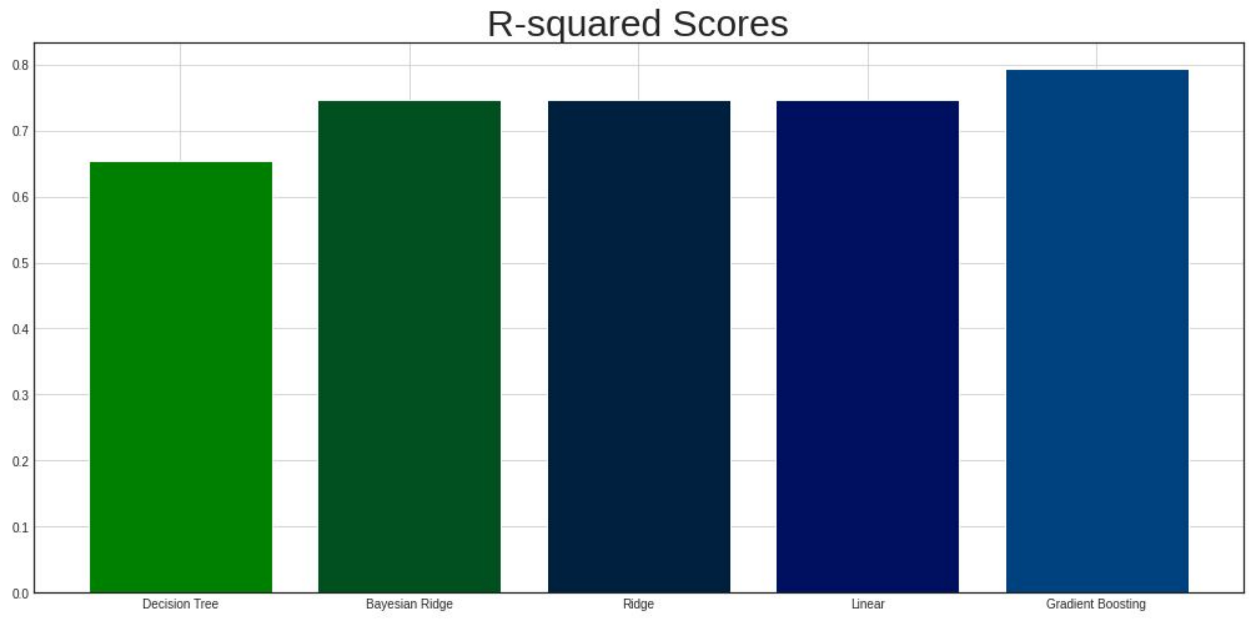


Figure 8: R-Squared Score Comparison

By using the correlation map and applying algorithms, a comparison between the variables contributing to happiness score is estimated. In this, out of all variables, health and income plays a very crucial role and has the most relative importance.

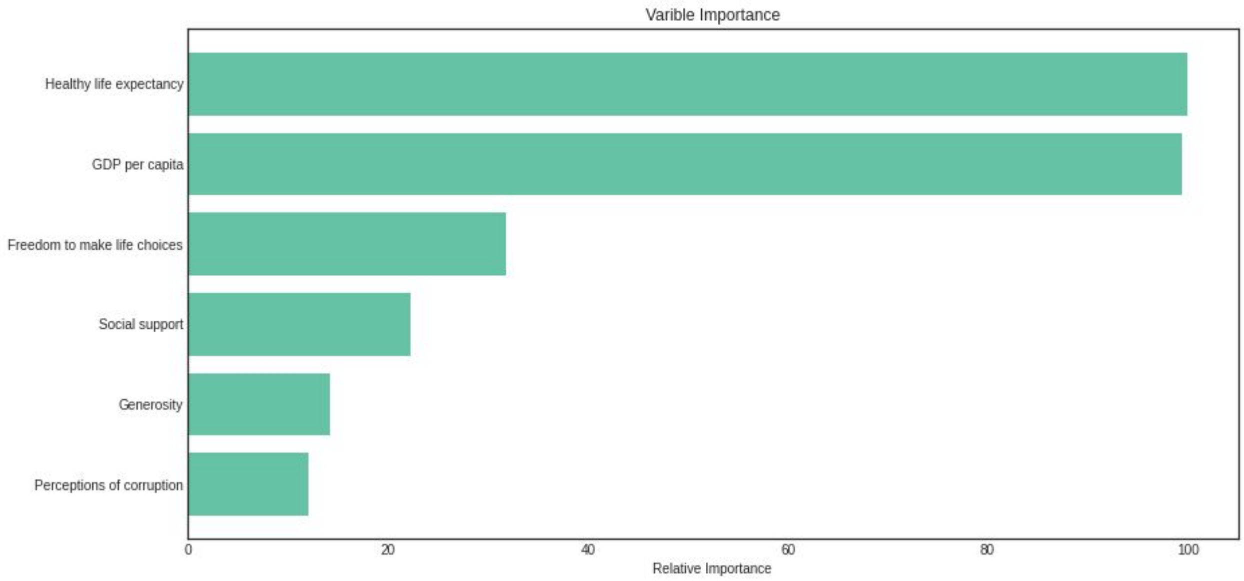


Figure 9: Variable Importance

- Conclusion

From the algorithms and graphs provided above, we can conclude that depending on the dataset selected, we find that Healthy Life Expectancy and GDP per capita contributed the most in the happiness score and are the most important variable in comparison with others contributing to happiness. Additionally, the most effective machine learning algorithm was found to be Gradient Boosting Regression. The model accuracy can still get increased , due to limitations we were not able to achieve better results. Limitations such as lack of resources. We tried adjusting the estimators and learning rate to understand differences in our models without managing to achieve any conclusive results.

- Future Scope

The project can further be continued considering all the countries latest survey and then comparing the differences and trends, providing the latest standing of that particular country in the world happiness ranking. The dataset used for this report is countrywise, there are more dataset available for cities as well, in depth analysis for cities can help the local government can help to improve the happiness score of that particular city. Additionally, we can provide a more content and comprehensive understanding by creating a dataset providing additional information.

- References

Cummins, R. A., Eckersley, R., Pallant, J., Van Vugt, J., & Misajon, R. (2003). Developing a national index of subjective well being: The Australian Unity Wellbeing Index.Social Indicators Research, 64(2),159-190.

Kohler, H. P., Behrman, J. R., & Skytthe, A. (2005). Partner?children=happiness? The effects of partnerships and fertility on well-being. Population and Development Review, 31, 407–445.

Jerome H. Friedman, (2002) Stochastic gradient boosting, Computational Statistics & Data Analysis, Volume 38, Issue 4, 367-378

Layard, R. (2005). Happiness: Lessons from a new science. New York, NY: Penguin.

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. Psychological Science, 7,186–189.

Ovaska, T.,& Takashima, R. (2006).Economic policy and the level of self-perceived well-being: An international comparison. The Journal of Socio-Economics, 35,308–325.

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Rid

ge_Regression.pdf. Accessed December 2020.