SDSC 6004 Group Project

World Happiness Score Report

Team: FANG WANG

Yuhong Fang 56373978 Zhe Wang 56537824

Contents

Abstract

Our team decided to write a world happiness report. This topic was chosen because we believe its purpose is to investigate and understand the scientific basis of happiness. The report introduces the globally available data on national happiness, and reviews relevant evidence from emerging happiness science, showing that people's quality of life can be assessed consistently, reliably and effectively through various subjective well-being measures, collectively referred to as then in subsequent reports Called "happiness". The World Happiness Report 2019 ranks 156 countries/regions by happiness level. This data is very important because our sense of happiness is affected by goal awareness and we feel that the things you do in life are worthwhile. Positive emotions are more important to us than no negative emotions, although both are important.

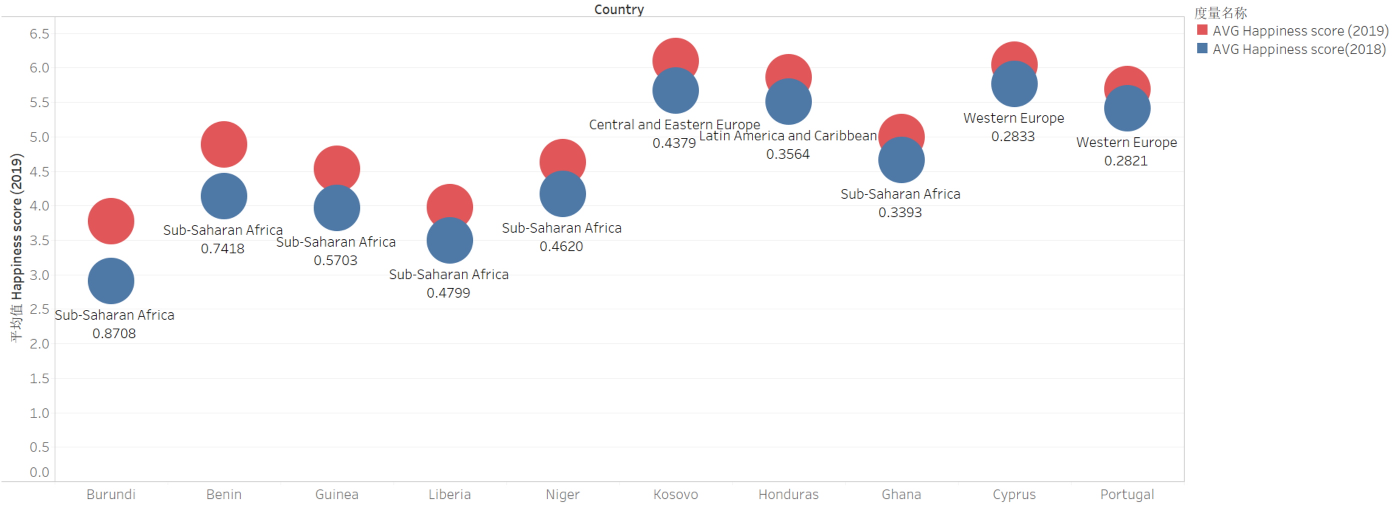
Our team's goal is to find out the main factors that caused the country's ranking or score to change between 2018 and 2019 reports, and the reasons for the country's annual changes. We also want to know which countries or regions rank the highest in overall happiness, and each of the six factors has an impact on happiness, and which country has significantly increased or decreased happiness. We will pay more attention to the target score, the happiness score. This data will allow us to analyze from thousands of people around the world and study ways to improve people's happiness. We look at the scores and explore the extent to which people have negative or positive emotional states in their daily lives. The most useful attributes of these data are GDP per capita, degree of social support, life expectancy, free choice, generosity, and degree of corruption. In short, current data may be fun and challenging, but our team will find the correlation between each variable.

Project Business

When conducting data analysis, we must consider various business issues. In general, we have several problems and goals, which we hope to solve through data analysis. First, we want to know what are the main factors of overall happiness? Secondly, we want to know which variables and models can be used to understand the happiness level of local people simply and directly. Our analysis adds value to the globalization trend of companies and companies that are seeking globalization or have already globalized.

When analyzing, we hope to narrow down which key variables contribute to overall well-being. We discovered this and found that if these key variables are significantly different between 2018 and 2019, this will result in a significant change in happiness levels. We found that our three key variables are health (life expectancy), social support and economy (GDP per capita). These variables contribute the most to the overall happiness of a country and have a positive correlation. We also noticed that there is a positive correlation between health and economy, with an R-squared of 0.72. As described below, in the descriptive analysis, we found that from 2018 to 2019, happiness increased and decreased significantly. These include Burundi, where the country's economy, freedom and health have improved, and happiness has increased significantly. Malaysia was also discussed. Malaysia also reduced happiness because the decline in dystopia was greater than the happiness score. Therefore, based on our analysis, we found that the most influential variables that affect happiness scores are health, economic and social support. If these variables change significantly in a country, the impact will be felt in society. We proved this when analyzing major changes in national happiness scores.

Statistics of Regions with Increased Happiness Scores

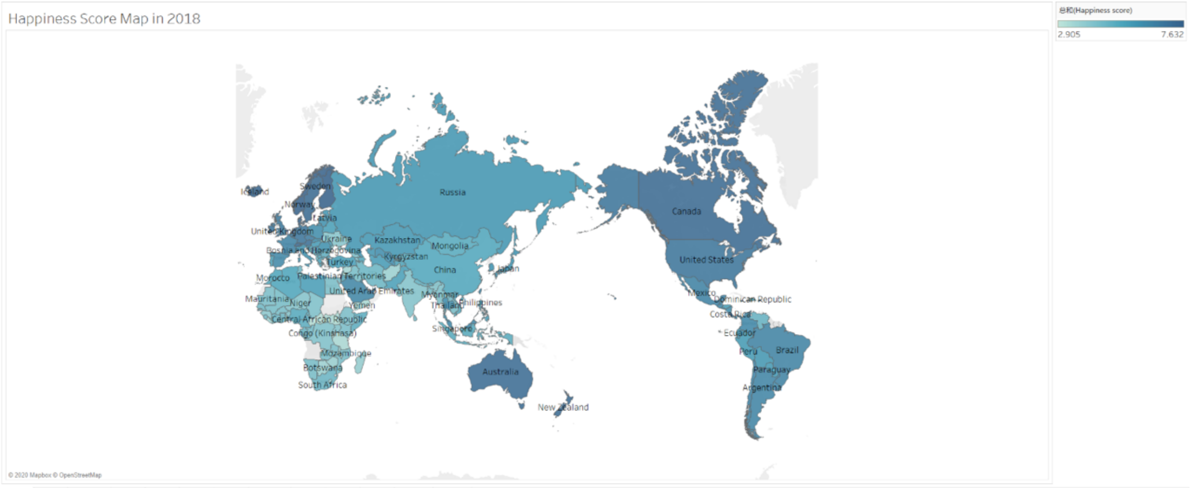


Data Understanding

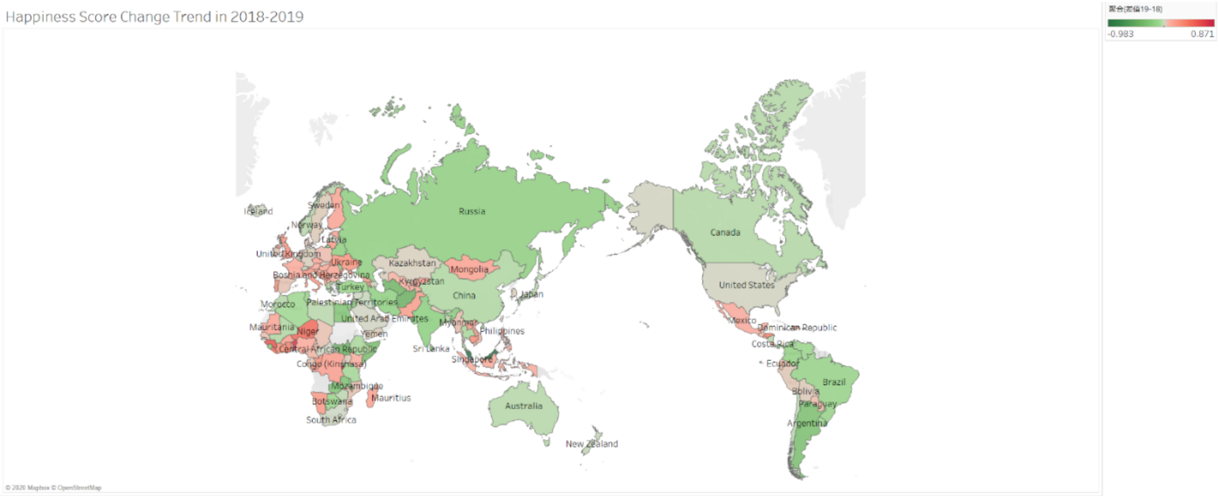
To understand our data, we must perform a simple analysis of our data set. The columns in this dataset include Country, Region, Happiness Scores, Ranking, Economy (GDP), Health, Freedom, Social Support, and Generous Dystopian Residuals. Our first variable is the country, which is a nominal variable. It contains 156 countries. Our second variable is Region, which is also a nominal value. This divides the data into regions by continent. Our third variable is Ranking, which is a number, and ranks the happiness of each country from 1 to 156 according to the world happiness score. We also obtained a numerical variable of GDP per capita, which allows us to gain insight into a country's economic situation. Health care is a digital variable based on the availability and quality of health care. Freedom is a numerical variable that reflects the degree of freedom in the area. Social support is a numerical variable that reflects the development of public assistance in the region. Dystopia: Anti-Utopia is used to correct situations where various indexes are rising but people do not feel happy, or when various indexes are falling but people feel happier.

Country	Happiness Score	Dystopia	GDP per Capital	Social Support	Healthy Life Expectancy	Freedom to Make Life Choices	Generosity	Perceptions of Corruption	Region	Rank
Finland	7.769	2.714	1.340	1.587	0.986	0.596	0.153	0.393	Western Europe	1
Denmark	7.600	2.393	1.383	1.573	0.996	0.592	0.252	0.410	Western Europe	2
Norway	7.554	2.241	1.488	1.582	1.028	0.603	0.271	0.341	Western Europe	3
Iceland	7.494	2.401	1.380	1.624	1.026	0.591	0.354	0.118	Western Europe	4
Netherlands	7.488	2.393	1.396	1.522	0.999	0.557	0.322	0.298	Western Europe	5
Switzerland	7.480	2.272	1.452	1.526	1.052	0.572	0.263	0.343	Western Europe	6
Sweden	7.343	2.246	1.387	1.487	1.009	0.574	0.267	0.373	Western Europe	7
New Zealand	7.307	2.127	1.303	1.557	1.026	0.585	0.330	0.380	North America and ANZ	8
Canada	7.278	2.193	1.365	1.505	1.039	0.584	0.285	0.308	North America and ANZ	9

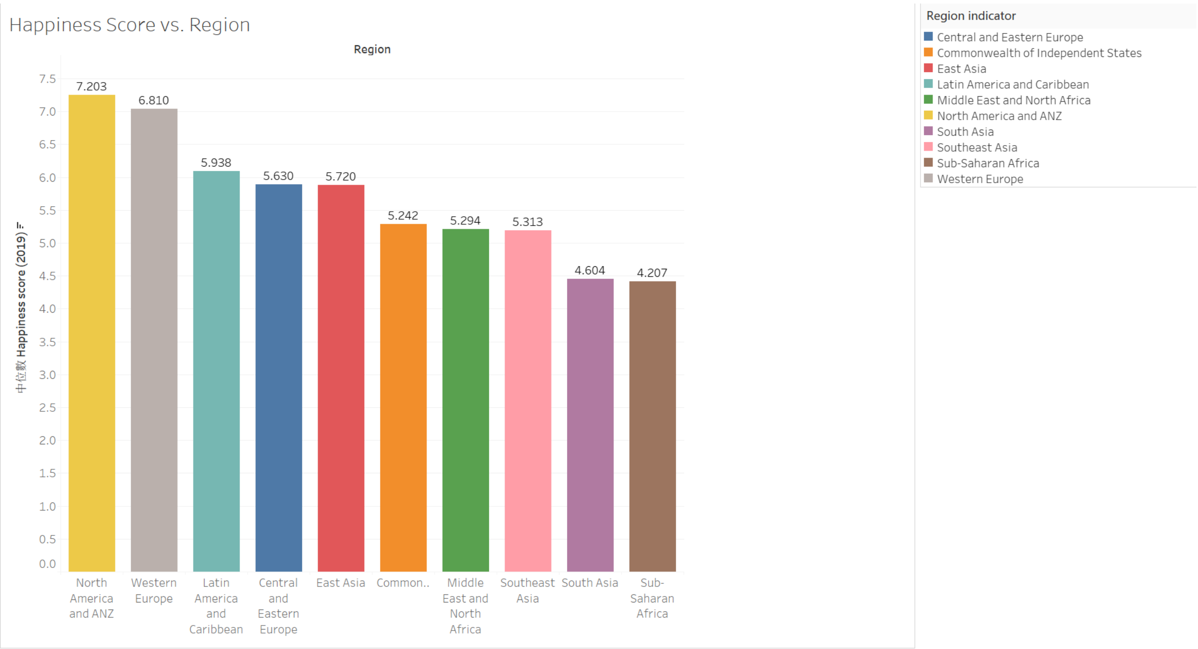
Data Visualization



Using Tableau to draw this happiness score map, we can find that Europe, North America, and Australia and New Zealand have more prominent happiness scores. These countries and regions are almost all developed countries. Relatively low happiness scores are concentrated in Africa and individual Southeast Asian regions. These countries and regions are developing countries, and the happiness scores of these countries and regions needs to be increased.



Through this map, we can clearly and intuitively see the changes in happiness scores in 2018-2019. Africa is showing a situation of increase and decrease, and the increase in China and Africa during the period is more obvious. The European region showed a small increase. In Southeast Asia, except for Malaysia, which saw a large decrease, other countries and regions remained basically stable. South America showed a slight downward trend.



By observing the map created by Tableau, we can find that high happiness score countries are centered on Western Europe and North America and ANZ, South Asia and Sub-Saharan Africa are relatively un-happiness. For detailed information, there are 13 countries from Western Europe and 4 countries from North America and ANZ in the top 20 most happiness countries list. In the top 20 least happiness countries list, 14 countries are from Sub-Saharan Africa.

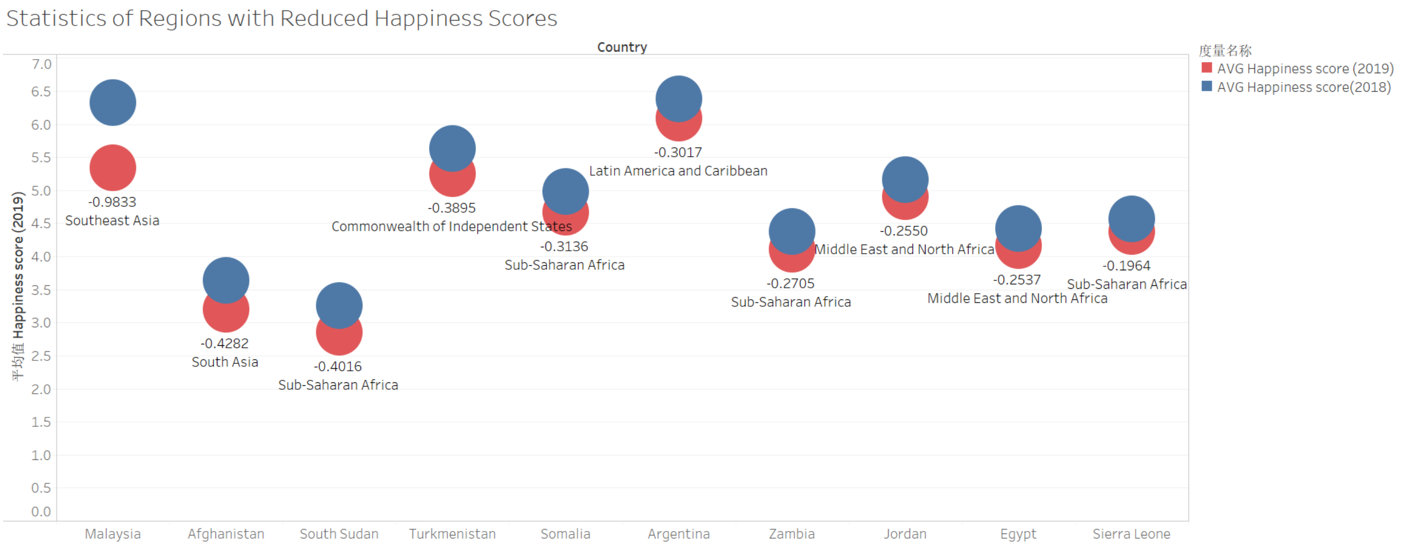
Descriptive Analytics

We get data from both 2018 and 2019 World Happiness Report and calculate the rank changed from 2018 to 2019 for each country. There are only three countries' rank changed more than 20, including Malaysia, Guinea, and Benin. Among them, Malaysia dropped the most for 45 ranks from 35 to 80 and Benin increase the most for 35 ranks from 136 to 102. Guinea increased 22 positions from 140 to 118.

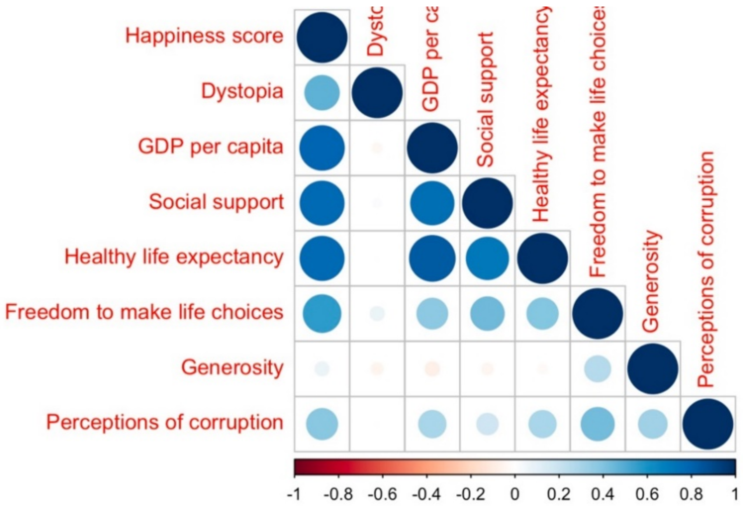
Malaysia

Year	Rank	Score	Dystopia	GDP	Social	Healthy	Freedom	Generosity	Corruption
2018	35	6.32	2.51	1.16	1.26	0.67	0.36	0.31	0.06
2019	80	5.34	1.33	1.22	1.17	0.83	0.51	0.26	0.02
Diff	-45	-0.98	-1.18	0.04	-0.09	0.16	0.15	-0.05	-0.04

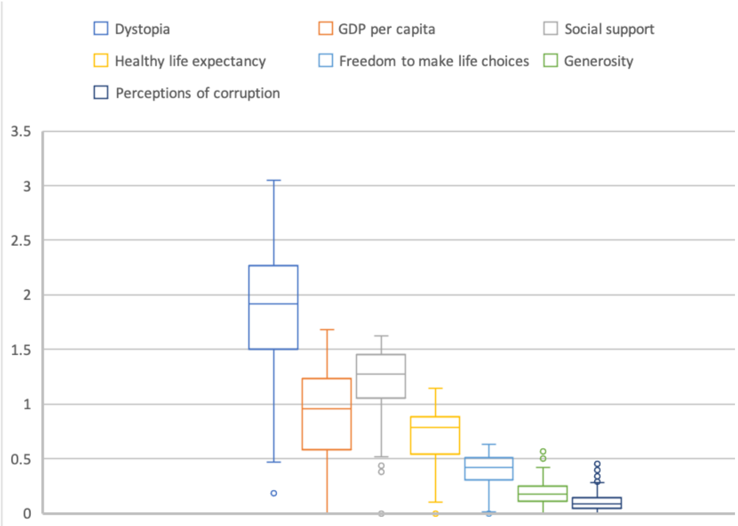
We are interested in what cause the decrease of Malaysia that huge in one year. By observing the data, we find that the dystopia has a larger decrease than happiness score. Which means, in general, the six factors increase slightly but people in Malaysia feel less happiness. Based on the data, we may conclude that people in Malaysia cared about the social support, generosity, and corruption than the global average, and they are more sensitive to the decrease of these features.



In top 10 rank decrease countries, 1 from Commonwealth of Independent States, 1 from Latin America and Caribbean, 1 from South Asia, 1 from Southeast Asia, 3 from Middle East and North Africa, and 3 from Sub-Saharan Africa. In top 10 rank increase countries, 1 from Commonwealth of Independent States, 1 from Latin America and Caribbean, 2 from Central and Eastern Europe, 1 from Southeast Asia, 1 from Western Europe, and 4 from Sub-Saharan Africa. We observe that Sub-Saharan Africa is an interesting region because it contains both 3 highly decreased countries and 4 highly increased countries. Also, this region contains many low happiness countries. So, this is a region with low happiness and not stable. By looking at the history and the nowadays situation, it is not hard to explain this observation. Maybe due to the bad economy and wars, their happiness is not able to increase hugely. Also because of the start or stop of the war or international economic assistance, their feeling of happiness changed rapidly.



To understand the features better, we provide the correlation plot to show the correlation between seven features and happiness score. The happiness score has positive correlation with all seven features because the data was made as this before published. It's correlations to GDP per capita, social support, and healthy life expectancy are all between 0.7 and 0.8. Although it is not high enough to be concluded as high correlated variables, it is still relatively high, and we can trust that there is some linear relationship between them. Also, GDP, social support and healthy have relatively high correlations between them. They may linearly be dependent to each other. It can be explained because usually a rich country with high GDP per capita can have enough money to build a good health system to extend people's life, also, they have money to do social support such as charity and disabled service. Because these features are all important information to describe a country, so we cannot simply combine them because of the correlations.



The happiness score is simply the summation of all seven attributes and all seven attributes were already managed before published and assigned weights to describe the happiness score. So, by observing the boxplot, it describes the distribution of seven features. Dystopia has the largest median and range, which means its weight is the largest and contains the most information. Perceptions of corruption's median and variance are both low, so it is relatively not important in the construction of happiness score.

Prediction Analysis

Happiness score is constructed as the summation of all seven attributes, so obviously there is a linear relationship between themselves and it is not interesting and not useful to analyze the relationship between happiness score and other attributes. We are interested in can we use less attributes or less information to predict the happiness situation of a country. How if we do not know that country a lot? Also, features may not maintain the linear relationship after the loss of information, so we interested in which model can predict the happiness level the best.

We separate each attribute into a three factors variable which contains low, normal, and high. They are separated based on the 1st quantile and 3rd quantile, values lower than 1st quantile are marked as low, and values higher than 3rd quantile are marked as high, all between are marked as normal. This is because we usually do not know a country's attributes in detail, especially use a number to describe it. Use low, normal, and high is much easier for people to illustrate a country's attributes.

Due to the linear relationship between happiness score and other variables, although data is changed from continuous to discretized, the linear relationship will lose partially but will not disappear instantly. So, we can imagine that we can only use several or even only one to get a good guess of happiness level. Because dystopia is a calculated variable and cannot provide by common sense, so it is totally useless here, so we simply remove it from this study, so we only have six attributes to construct the happiness level. We use Weka with 10-fold cross validation to provide the classification result and use the accuracy, mean ROC area and Kappa value as the output to compare the models. Because the data contains the linear relationship and all are converted into categorical data, and the simple structure, we expect to see that simple models can predict the data well instead of using some complex machine learning models and tree models and logistic models can have relatively good performance.

All Six Variables:

Model	Accuracy (%)	Mean AUC	Kappa
J48	77.56	0.799	0.6325
Naïve Bayes	76.92	0.862	0.6269
KNN	75.64	0.839	0.6021
Random Forest	74.36	0.867	0.5960
SimpleLogistic	73.08	0.824	0.5692

The highest accuracy is 77.56% provided by J48 tree model. It is not high enough to say this is a good prediction model for this case. By comparing the result between these five models, J48 tree model has the highest value in accuracy and Kappa but its mean AUC is lower than Naïve Bayes. We cannot simply conclude any of them as the best model, so we conclude that J48 tree model and Naïve Bayes model are the two-best model in this case.

We use many variables to predict the happiness score, but the accuracy is not good enough. So, we think that can we use less features to predict the happiness score level but maintain a similar accuracy? By previous observing and analysis, we find that GDP per capita, social support, and healthy life expectancy are three variables that have relatively high relationship with happiness score. So, we expect to see that we can still get a good happiness score level prediction by only using these three variables.

Three variables: GDP per Capita, Social Support, Healthy Life Expectancy

Model	Accuracy (%)	Mean AUC	Kappa
J48	78.85	0.808	0.6499
SVM	78.21	0.834	0.6458
Random Forest	76.28	0.858	0.6176
KNN	75	0.827	0.5948
SimpleLogistic	75	0.826	0.5948

By removing three variables, the accuracy increased slightly which represents that the three variables that were removed lead an overfitting and caused the result of prediction bad. J48 and SVM provide two best result. Although the accuracy is still not higher than 80% even 90%, the decrease of number of variables make the model easier so the result is more acceptable.

Because these three variables have a positive correlation to each other, so we are interested in can we use only one of these three variables to predict the happiness score level and still get an acceptable prediction result?

By only using GDP per capita to predict the happiness score level, all models provide similar result with accuracy equals to 74.36%, AUC around 0.765, and Kappa is 0.5897. Although the accuracy decreases for more than 4 percent, the result is still acceptable because we only used one variable. By only using social support, the accuracy is 67.95%, and the accuracy is 70.51% by only using healthy life expectancy. The accuracy is lower than by only using GDP per capita so may be use the economy level to predict the happiness level for a country is the best if we only want to use one variable to predict.

Evaluation

We were careful to avoid over fitting and other negative influences through our use of information gain ratio and carefully removing and combining certain attributes such as happiness rank to get a clear reading on the relationship amongst our most and least informative variables. We further compared our most informative and least informative variables to determine the coefficient of determination between each variable. This allows us to accurately evaluate our results.

We try to use a variety of models, so that we have a better understanding of how to evaluate and later implementation and deployment of discoveries. Since this data set is an abstract concept, we considered all the different business areas to which the data set applies and selected two areas that will ultimately affect all other areas: globalization and government.

We understand that many different companies have expanded their business into new areas around the world. Globalization brings new opportunities for success, and through the evaluation of the data set, we believe that we can help any company expand its business. Our evaluation of the data set allows us to understand the relationship between the most important factors of happiness and the relationship between them. This understanding of happiness factors will make it possible for companies to decide that they want to deploy businesses in emerging regions.

We believe that deploying business in new regions according to the World Happiness Report will help increase the company's future profitability. By carefully analyzing our reports, companies can choose where to deploy new businesses based on a country’s overall happiness score. For example, if a clean energy company wants to deploy a new industry business but cannot decide which country/region it is in, they can use the model we created to determine it based on the country’s economic status, population, social support, government corruption and health status New location.

Compare with Published Result

Through this data set, we have discovered many relationships between happiness and information variables. We found that as certain variables (such as Health, Economic and Social Support) increase, the overall ranking of happiness will increase. This is substantial because it allows us to understand happiness from a macro perspective. When processing and drawing certain relationship diagrams, we realized that there is indeed a linear relationship with these information variables. Our analysis of Malaysia can prove this point. Due to the reduction of other factors such as health, economy and government trust, their happiness scores have dropped significantly. This illustrates the relationship between happiness. In a few areas, happiness scores increased significantly due to recovery after the war. The medical and social support of the recovered country gradually picked up. However, we have noticed that as we approach the top rankings and the gap between countries in some informational variables becomes larger and larger, some smaller variables become differentiating factors.

This was an important discovery as it leads us to understanding how the top countries were ranked. We noticed that while some of the attributes such as happiness score, economy, and health were very similar among countries, usually within a tenth of a difference from each other, other countries were leading in some areas substantially, particularly the social areas such as generosity, social support, and freedom.

Deployment

There are many ways to utilize and deploy our data set. In the deployment analysis, we decided to segment it by who will use this information and how they will use it.

We will determine which organizations or companies will use our analysis results. How can these identified companies or organizations use our analysis results to help them? We can discuss how they use this information.

First, national think tanks can use these results for a given country/region to point out their focus on pushing the country to its peak. For example, our analysis has determined why each country’s happiness index is low or high. Similarly, why give this country this happiness score. Therefore, national think tanks can discuss their policies based on our analysis to improve the country’s lack of happiness. More specifically, they can use our descriptive analysis to view their health and GDP. We have seen that the R-squared between health and GDP is 0.75 and there is a positive correlation. Therefore, if government officials find that there is room for improvement in their healthcare based on our descriptive analysis, they can provide a policy to increase investment in healthcare for their country. They can use our analysis to prove that investing in healthcare is a wise choice, because our descriptive analysis is positively correlated, their GDP per capita (per capita) will also increase, and the overall happiness of citizens will also increase. Therefore, based on our happiness analysis, government officials can know which country they should pay more attention to. This will help government officials to provide a better life for their citizens. Government officials will not only find this information very valuable, but also change the focus of development for a period in the future to better serve citizens. In addition, medical R&D companies can use this report to lay out relevant industrial chains in economically developed and health-conscious areas. This will bring more advanced medical technology and advanced medical concepts to local people. Having a positive environment is more conducive to product research and development. Both financial support and policy support will bring benefits to enterprises. From this data set, a more in-depth investigation of the situation in developing countries can also be made. For some emerging economies, although the health variables are not optimistic, the economic variables have increased rapidly, and there is more room for growth in the later development. This is also a direction that enterprises should pay attention to. At the same time, the level of government corruption is also very important. Enterprises do not want to develop in areas with poor business environment. A lot of practical experience shows that places with high levels of corruption are not conducive to economic development. According to the data set, happiness scores are also reflected. This also has a fundamental impact on the development of enterprises.

Finally, other industries can also deploy our data sets, such as new energy, clean energy, fast-moving consumer goods that favor health, and the Internet and other industries can use this data set to adjust and develop the business of enterprises. It is more in line with the development trend of smart cities.

In a word, the yearning and pursuit of happiness for residents of every country remains unchanged, all to make life better. Through the happiness score report, more industries, governments and organizations can make decisions about the future development of the city, have a better judgment on the future development, and timely adjust the direction of future development. Through a series of analyses, the government can make citizens’ lives better, have a higher happiness index, people’s healthy life span, and the economy can be more stable. Companies can avoid losses through analysis, find profit growth points, better serve the city, create tax revenue, and grow steadily to make employees' happiness index higher. The happiness index is always closely related to everyone and the entire society. It will always serve the smart city.