Scientometric and Psychological Analysis of World Happiness

1. Introduction

Happiness is a positive emotional state, and it can be further extended to many concepts like life satisfaction, psychological well-being, flourishing and so on. Measuring the nation's happiness score allows us to understand the country's social climate and the reasons behind nation's happiness level.

The objective of this report is separated into three parts. First, to analyze the correlation between the happiness score and multiple socioeconomical factors by using the data of 149 countries. Second, to predict the happiness score according to these factors. Third, to determine significant factors contributing to the overall happiness score.

In order to analyze the relationship between distinct factors, both machine learning methods, such as Decision Tree, Random Forest, XGBoost, and deep learning methods, such as Multilayer Perception (MLP) and Convolutional Neural Network (CNN), are applied for regressing the happiness score index. Furthermore, hyperparameter tuning is provided to improve the accuracy of the regression. Error analysis is complemented to compare the performance of different models. Finally, psychological theories and research are applied to support our results of analyzing such correlation between significant factors.

2. Data Source

The world happiness data is collected from Kaggle. There are two separate datasets. In the first dataset, a total of 149 countries' entry is recorded in 2021, with 7 factors, including ladder score, logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. Another dataset has a similar structure to the first dataset, which was recorded between 2005 and 2020.

Therefore, in order to have a better performance in machine leaning and deep learning, two datasets are combined into one for establishing a larger sample size to predict the happiness score index. The data between 2005 and 2020 is divided into our training set, while the data of 2021 is divided into our test set.

In addition, some repeated factors, which are just scaled to the original data, are deleted, including "Explained by: Log GDP per capita", "Explained by: Social support", "Explained by: Healthy life expectancy", "Explained by: Freedom to make life choices", "Explained by: Generosity", "Explained by: Generosity", "Explained by: Perceptions of corruption". Besides, some factors are different in two datasets, so those unmatched factors are also deleted.

Inspired by Prathamesh Dinkar, Random Forest regression is applied to find out the important feature of happiness score and world maps are produced for visualizing the data.

2.1 Structure of the Dataset

The definition of the variables is as follows:

Variables	Definition
Ladder score	The happiness score
Logged GDP per capita	The logged value of economic output of a nation per person
Social Support	Perception of caring from surrounded people, binary scores of 0 and 1 of the question: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
Health life expectancy	Number of years in a healthy life stage
Freedom to make life choices	Binary scores of 0 and 1 of the question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
Generosity	The residual of regressing the national average of responses to the question: "Have you donated money to a charity in the past month?" on GDP per capita
Perceptions of corruption	Binary scores of 0 and 1 of the questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?"

The summary statistics of each variable is as shown:

Variables	Mean	SD	Minimum	Maximum	No. of missing data
Ladder Score	5.469	1.115	2.375	8.109	0
Logged GDP per capita	9.372	1.145	6.635	11.648	36
Social Support	0.813	0.118	0.290	0.987	13
Health life expectancy	63.472	7.370	32.30	77.10	55
Freedom to make life choices	0.746	0.140	0.258	0.985	32
Generosity	-0.001	0.158	-0.335	0.698	89
Perceptions of corruption	0.746	0.181	0.035	0.983	110

3. Data Cleaning and Visualization

From the above summary statistics, it is found that missing data exists in all explanatory variables. Therefore, the missing values are replaced by their corresponding means.

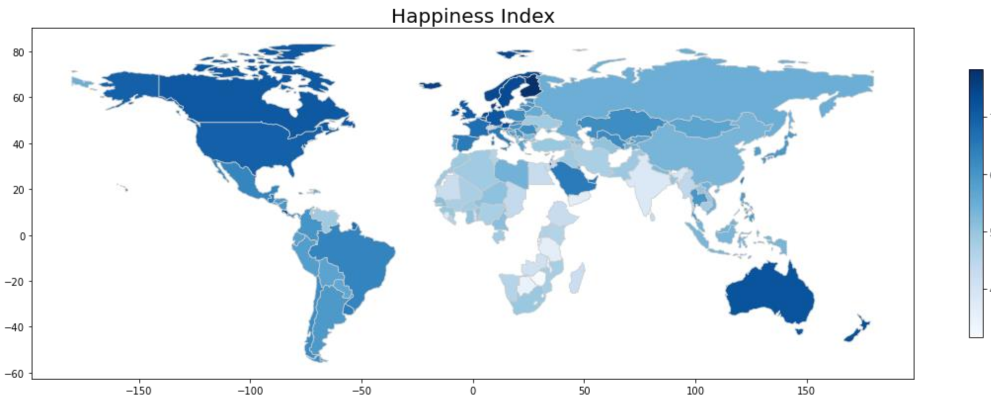
For a better understanding of each variable, we plot the histograms of distinct variables. According to appendix B, the ladder score shows a mostly symmetrical distribution. Apart from the ladder score and generosity, all other variables show left skewing. Generosity is the only variable, which shows left skewing.

Figure I Correlation Heatmap

From the correlation heatmap above, some factors are strongly correlated to the ladder score, such as logged GDP per capita (r = +0.78), healthy life expectancy (r = +0.74), and social support (r = +0.71). This implies that these factors have a positive effect on the nation's happiness score, whereas freedom to make life choices (r = +0.53) and perceptions of corruption (r = -0.42) have a medium effect, and generosity (r = +0.17) has a small correlation toward happiness score.

As for the boxplots of different variables, according to appendix C, although some extreme values exist in some of the variables, the happiness index and other observed data fluctuate with each country. It is possible that exception stats exist in some countries. Since all data lies on their reasonable range, these extreme values are not considered as outliers.

Figure II World Map with Happiness Index



By the world map with their corresponding happiness index, it is observed that North America, Australia, and Europe obtain an outstanding happiness index, when Africa and South Asia obtain a relatively low happiness index.

4. Method/Analysis

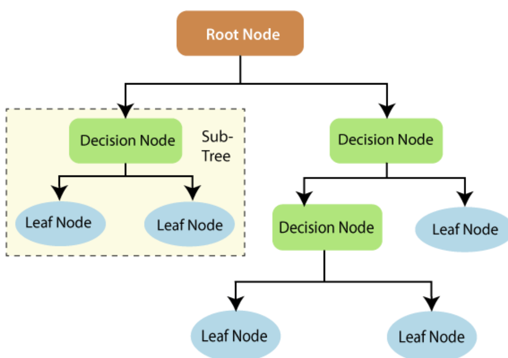
4.1 Machine Learning

4.1.1 Decision Tree

Decision Tree is a popular supervised machine learning algorithm for dealing with regression and classification problems. This algorithm will train up a decision rule by the training dataset, then use the decision rule to predict the value or classify the class of target.

The decision tree is a tree-like model, it uses three major nodes to generate the entire model: root node, decision node and leaf node. The model will start from the root node, then further divide into different decision nodes. If the node stops splitting, it is a leaf node. Each subsection of the whole tree is called sub-tree.

Figure III Standard Decision Tree



In this project, we would like to use a regression tree to predict the happiness score, instead of a classification tree. In the case of regression, the classification and regression tree algorithm (one of the popular algorithms to generate a decision tree) would like to split the data in the tree by minimizing the loss function, such as a residual sum of squares. The optimal decision rule is the rule which has the last residual sum of squares.

Decision Tree is a powerful algorithm, but it still has a high chance of overfitting the problem. Overfitting means the model is too complex, having large variance and small bias. The model can fit into the training dataset very well, but has small prediction power in the testing dataset, which is an unfavorable result. Therefore, Random Forest is postponed solving the problem.

4.1.2 Random Forest

Random Forest is an ensemble algorithm, it generates a bunch of Decision Trees to build a whole model for regression or classification tasks. For regression problems, the prediction value is the mean of the prediction value of different Decision Trees. For the classification problem, the predicted class is voted by each Decision Tree.

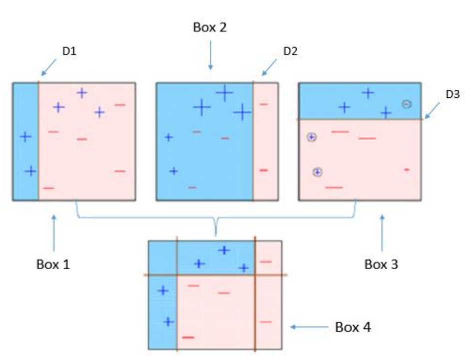
The concept of bagging and boosting is used in Random Forest since the model has created a different Decision Tree by one dataset. Bagging is the process of drawing a sub-sample. It picks up a subsample randomly and creates a Decision Tree with the subsample. After creating a tree, the subsample will be put back into the training dataset and repeat the above process unstill the last Decision Tree is created. Since each sub-sample draws from the same training dataset, those sub-samples may overlap but each Decision Tree still varies. Meanwhile, the concept of boosting is similar to bagging but boosting combines the tree, which has less prediction power into a stronger tree. Also, it adjusts the weight of the input that increases the weight of incorrect information in the input. It can allow a new tree to learn from the previous result to have stronger prediction power. Applying the concept of bagging and boosting can prevent the problem of overfitting.

4.1.3 XGBoost

XGBoost is a supervised learning algorithm, which is implemented of the gradient boosting trees algorithm for dealing with tabular data problems, by using regression and classification trees.

Gradient boosting is also an ensemble algorithm, constructed from Decision Tree. The idea of boosting is combining a set of weak learners and to continuously improve the weak learners and integrating them to form a strong leaner. Different from bagging, instead of generating a set of independent trees randomly, each successive tree compensates for weakness by correcting the prediction errors in and are dependent on prior models.

Figure IV Standard Gradient Boosting



XGBoost model is fitted by combining the loss function and penalty function for the complexity of the model. The value of losses can be used as the basis for judging the quality of model training, which are the errors between predicted and target values. Since the endless complication may lead to problems of over-fitting, the penalty function, regularization, is needed to avoid that. By minimizing this objective function after adding new models, the optimal XGBoost model is constructed.

4.2 Deep Learning

4.2.1 Multilayer perceptron (MLP)

Multilayer perceptron is a feedforward artificial neural network for regression and classification problems by changing the output of the model. For regression problems, the model should have a linear output. At the same time, the model should have the activated output for classification problems.

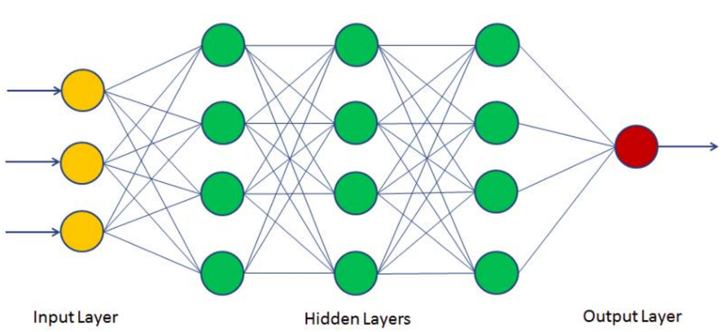
Multilayer perceptron is an algorithm for solving different problems by simulating the architecture of the human brain. There are a lot of neurons in our brain, and thanks to these neurons humans can learn, think and memorize a thing. Multilayer perceptron would like to reproduce the function of those powerful neurons in the mathematical model.

Multilayer perceptron is three main layers to build the entire model, including the input layer, hidden layer, and output layer. In the input layer, the model will receive the data with different features and transform the data into a hidden layer. The hidden layer contains a lot of neurons, that neuron is the linear combination of weighting and bias of the input data. After all the data is activated by activation functions, such as Tanh, ReLU, or Sigmoid function, those activated data will be the input of the next hidden layer. Finally, the model will get output by the linear combination of the neuron in the last hidden layer. The formula of each neuron is

$$Z_j = \omega_1 x_1 + \omega_2 x_2 + \cdots \omega_k x_k + b_j$$

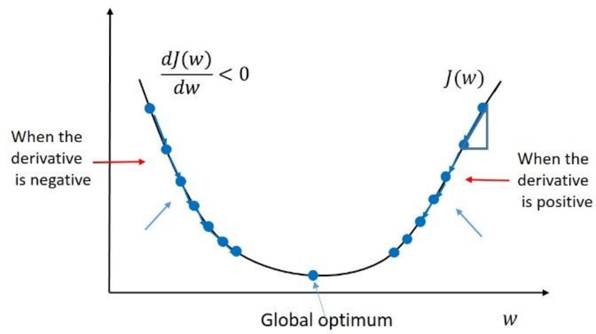
where x is the input data, b is bias term.

Figure V Standard Multilayer perceptron



Since the model cannot get an optimal result by processing the feedforward at once, after getting the result the model will perform Backpropagation. Backpropagation allows the model to adjust weighting and bias in each linear combination in order to minimize the loss function. The process of Backpropagation is achieved by Gradient Descent, which is an optimization algorithm by using the negative gradient to update the parameter. In the context of deep learning, the model updates the weighting and bias by the Gradient of loss function, which is the mean square error function. The parameter will be updated from the output layer to the first hidden layer.

Figure VI Standard Gradient Descent



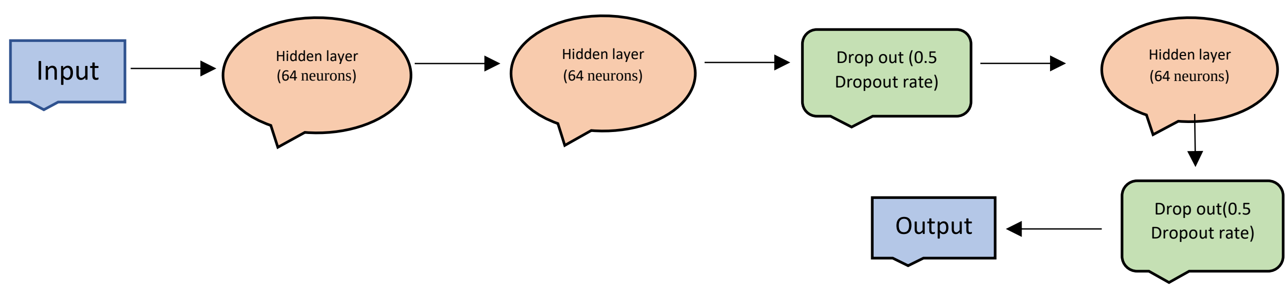
In this project, one input layer, three hidden layers and one final output layer are applied. Each hidden layer contains 64 neurons and uses ReLu function as activation function:

$$f(x) = \max(0, x)$$

There are three main advantages of ReLu function. One is that it can solve the vanishing gradient problem. For the vanishing gradient problem, many gradients may be exceedingly small which may abort the update of the parameter. The ReLu function can stabilize the convergence to prevent a vanishing gradient. Also, the ReLu is a simple mathematics function so it can reduce the computation time of the entire model. Besides ReLu is sparsity, many inputs will activate to zero. It reduces the complexity of the model and the variance that can avoid the problem of overfitting.

In the last two hidden layers, we would like to drop out some neurons by Bernoulli distribution. Because the regularization effect is performed by drop-out neurons, it forces the model to a less complex model to prevent overfitting.

Figure VII Architecture of Multilayer perceptron in this project

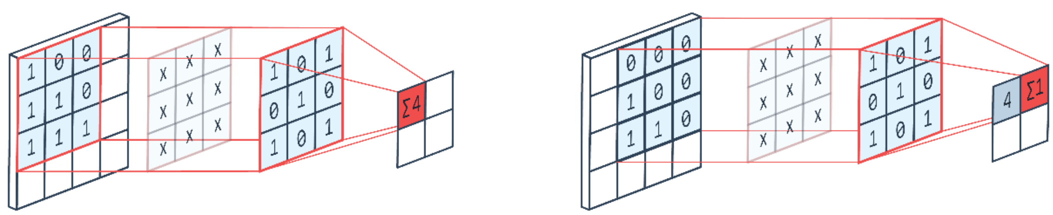


4.2.2 Convolutional Neural Network (CNN)

Convolutional Neural Network is another feedforward artificial neural network. The architecture of Convolutional Neural Network contains three major layers: Convolutional layer, Pooling layer and Fully connected layer, which like MLP simulates the human brain, CNN simulates the human visual system. The Convolutional and Pooling layer are like in visual cortex in human brain, trying to detect information outside. Usually, CNN is used for computer visual systems like facial recognition or image classification. However, in recent years the use of CNN has been extended to other fields, such as stock price prediction.

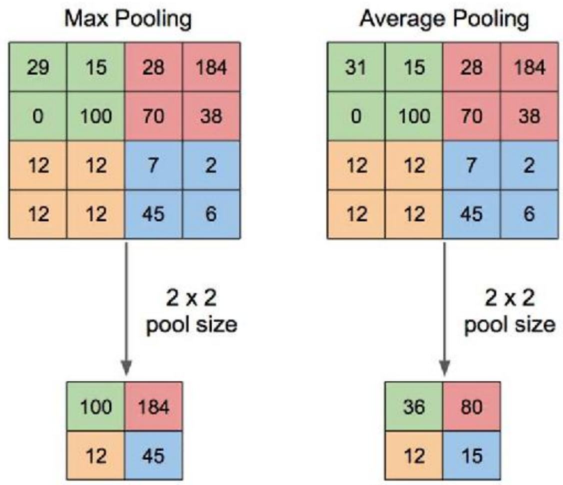
The Fully connected layer is just an MLP mentioned in the last section. Convolutional layer is a process of extracting some features or essential information from the input data by the filter. For example: If the input data is an image, Convolutional layer can detect the pattern or the major color in the image. In Convolutional layer, a filter moves across the whole input map vertically and horizontally. The filter convolution every single frame of the input map, after that we got an output map and treat it as an input map for pooling layer.

Figure VIII Convolutional layer



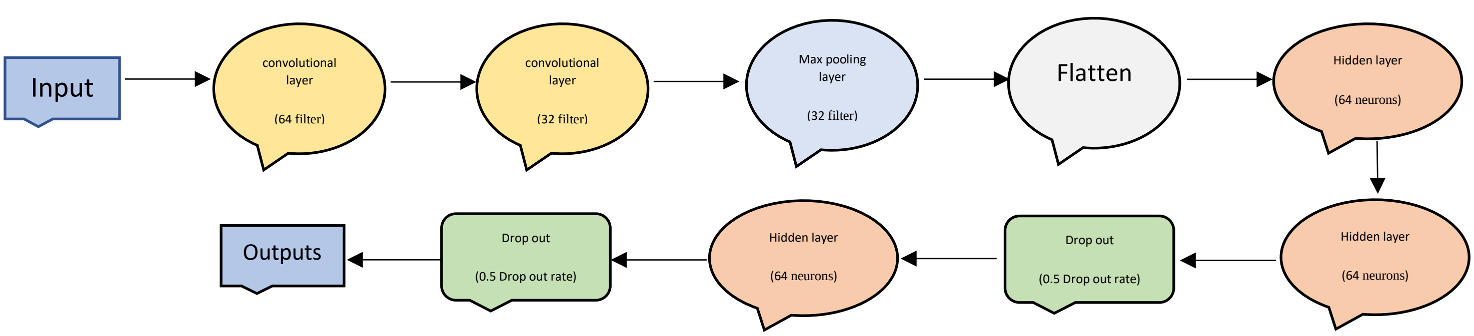
To reduce the complexity of the whole model, Pooling layer is added after convolution layer. Another filter is used for pooling the information in the output map of the last Convolutional layer. The filter moves the entire output map like the previous, but the filter just extracts the maximum value (Maximum pooling) or gets the average value (Average pooling) of the frame.

Figure IX Maximum pooling and Average pooling



In this project, 1-dimensional data is used, instead of 2-dimensional data like image. So, we use 1-dimensional convolutional neural network. Using two 1-dimensional convolutional layers with 64 filter and 32 filter respectively. After extracting the feature of the input map, we use the maximum pooling layer to reduce the complexity of the model. Then connect to a fully connected layer with three hidden layers and two drop-out layers.

Figure X Architecture of Convolutional Neural Network in this project



5. Results

5.1 Hyperparameter Tuning

Hyperparameter is the parameter, which controls the architecture and training process of the learning algorithm. Different combinations of hyperparameters can lead to large different results. However, the hyperparameters are not optimized in the training process, the model will not return optimal hyperparameters after fitting into the training dataset. Therefore, tuning hyperparameters is required for the best result.

In this project, we would use random search to tune the optimal hyperparameter. In the process of random search, we get a range of hyperparameters, and the model will evaluate combinations of hyperparameters randomly and find out the optimal hyperparameters by repeating the above process.

Below is the hyperparameter used in this project:

Machine learning	Deep learning
Decision Tree • Maximum depth = 5 • Minimum samples split = 82 • Maximum leaf node = 280	**MLP** • Optimizer = Adagrad • Epochs = 190 • Batch size = 77
Random forest • Maximum features = 2 • Minimum samples split = 15 • N estimators = 280	**CNN** • Optimizer = Sgd • Epochs = 277 • Batch size = 146
XGBoost • Subsample = 0.7 • N estimators = 500 • Minimum child weight = 1 • Maximum depth = 6 • Learning rate = 0.01 • Column sample by tree = 0.6 • Column sample by level = 0.799999999999 • seed = 10	

5.2 Evaluate

After training our model by the historical data, we would like to forecast the happiness index in 2021. To evaluate the performance of different models, we would use Mean Square Error after inverse the standard scaler. The formula and results are as below:

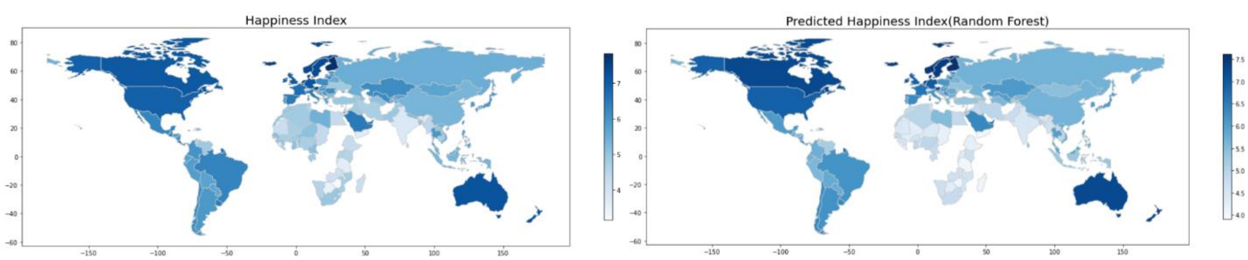
$$\varepsilon = \sum_{i=1}^{n} (prediction - true\ value\ )^2$$

	Decision Tree	Random Forest	XGBoost	MLP	CNN
MSE	0.240	0.118	0.140	0.258	0.211

The results show that both machine learning and deep learning models work well for regressing the happiness index, but machine learning models perform slightly better than deep learning models. Among the machine learning models, Random Forest has the best performance with the smallest mean square error.

Also, we can compare the actual Happiness index and predicted Happiness index by world map. We display the result of the regression of Random Forest below, since it got the smallest MSE among all five models. The results of other models are attached in appendix D. It is observed that our predicted happiness index in 2021 is close to reality.

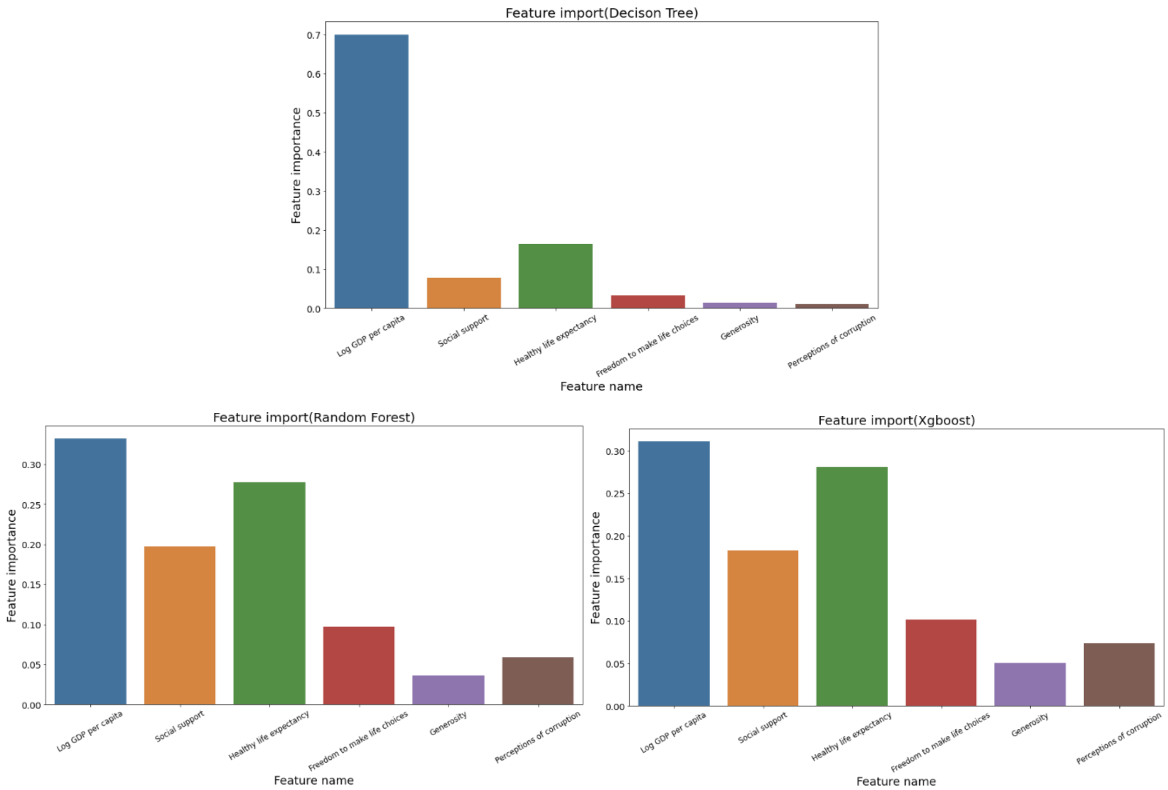
Figure XI Real Happiness Index and Predicted Happiness Index



5.3 Feature important

After fitting the data into our models and evaluating different models by MSE, we would like to find the feature which is more important to the happiness index than other features. Therefore, we use the function of "feature important" to find out the most significant feature. Note that this function only supports Decision Tree, Random Forest and XGBoost.

Figure XII Feature important of different models



It is observed in all models that logged GDP per capita has the greatest influence on the happiness index. Healthy life expectancy and social support have the second and third greatest influence on the happiness index. It is intriguing that freedom and the perceptions of corruption do not affect the happiness index very much.

6. Discussion

It is found that GDP per capita, healthy life expectancy and social support are the three most important features to happiness, while this result can also be supported by other theories and findings.

The significance of GDP per capita can be explained from an economic psychology aspect. Materialism, meaning the tendency to treat material desires as more valuable than spiritual matters, could be the reason explaining the phenomenon (Richins & Rudmin , 1984). When people own more possessions, their material desires would be satisfied. Therefore, national GDP and happiness are highly and positively correlated, an increase in national GDP would improve people's happiness or quality of life (Easterlin, 2012).

It should be noted that self-evaluation of one's health plays a more important role than objective measures (Britannica, 2021). Yet, there is no definite explanation between the relationship of happiness and healthy life expectancy. They may have mutual influences, as an unhappy emotional state would cause stress in the nervous system, while an optimistic person would tend to think positively that he or she would be healthy and live longer (Thompson, 2019). It is undeniable that there is a strong association between the two, but healthy life expectancy might not be a factor of happiness.

It is not surprising to see that social support is one of the most important factors. According to a psychology theory, people are born with an innate need to have stable interpersonal relationships (Baumeister & Leary, 2017). Social connections can greatly affect our emotions. A previous study also showed that happier people have stronger ties to their family members and friends (Diener & Seligman, 2002). Therefore, positive social support would make us happy.

There is a positive correlation between freedom to make life choices and happiness, yet, freedom contributes less to happiness compared with other features. One possible explanation is that the pattern of correlation varies across different countries. Freedom is more strongly correlated to happiness in wealthy countries, and weakly correlated in less developed countries (Brulé & Veenhoven, 2014). So, the overall association would be weaker when viewed as a whole picture.

7. Conclusion

With both methodologies of machine learning and deep learning, the world happiness index is well predicted within an exceedingly low error. Random Forest is the most optimal model for regressing the happiness index. By the analysis of correlation and important features, it is found that GDP per capita, healthy life expectancy and social support have the greatest influence on the happiness index, and most of the factors are correlated to the happiness index. The surprising results are observed that the freedom index of the countries only has a slight impact on the happiness index.

However, as for the limitation, since the computation power is limited by the specs of the computer, it is not able to tune out the most optimal hyperparameters of models, which can be improved by strengthening the specs of the computer or parallel computing. In our study, machine learning performed better compared with deep learning. It is believed that deep learning will perform better when having a high-end infrastructure and a larger sample size.

This study suggested that in order to improve citizens' happiness, countries should pay more efforts on the important features as their first priorities, such as economic development, healthcare services and social support. After that, since there is positive correlation between happiness index and freedom index, it is suggested to improve the human freedom to provide a fully pleasant environment for the citizens. The indicator of happiness index may change, when the overall living standards are improved. The best way to maintain a high happiness index is to keep ameliorating distinct aspects of life. We are looking forward to the improvement, that the worldwide are living in the pleasant and benevolent nation.

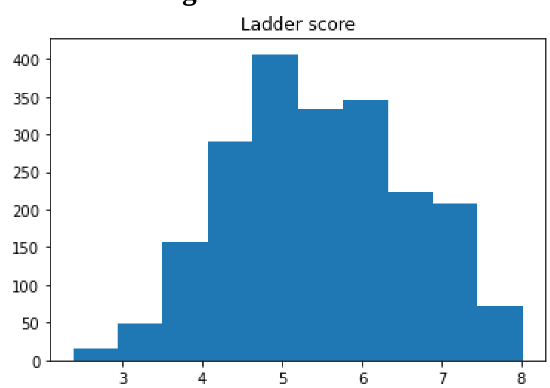
8. Appendix

Appendix A Summary statistics

	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000
mean	5.468979	9.371622	0.812637	63.471678	0.745989	-0.001062	0.745715
std	1.114936	1.144668	0.117804	7.369823	0.139662	0.157869	0.181236
min	2.375000	6.635000	0.290000	32.300000	0.258000	-0.335000	0.035000
25%	4.650750	8.480750	0.750000	58.992000	0.653750	-0.109000	0.698000
50%	5.391000	9.449000	0.833500	65.100000	0.763500	-0.017000	0.793000
75%	6.281500	10.343250	0.905000	68.525000	0.858000	0.084000	0.866000
max	8.019000	11.648000	0.987000	77.100000	0.985000	0.698000	0.983000

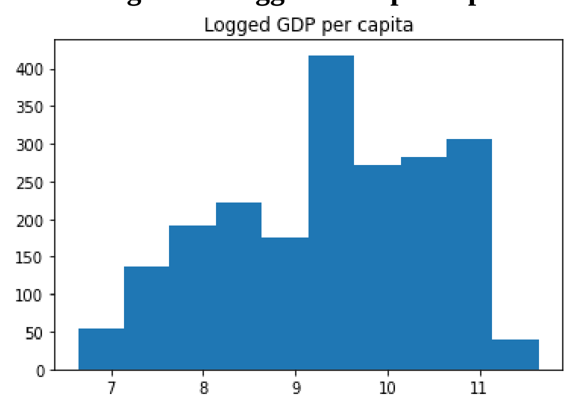
Appendix B.1

Histogram of ladder score



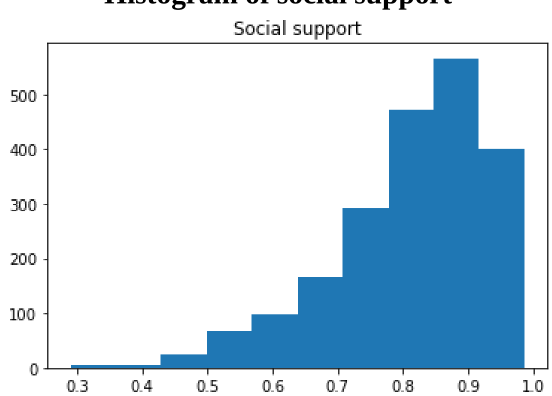
Appendix B.2

Histogram of logged GDP per capita



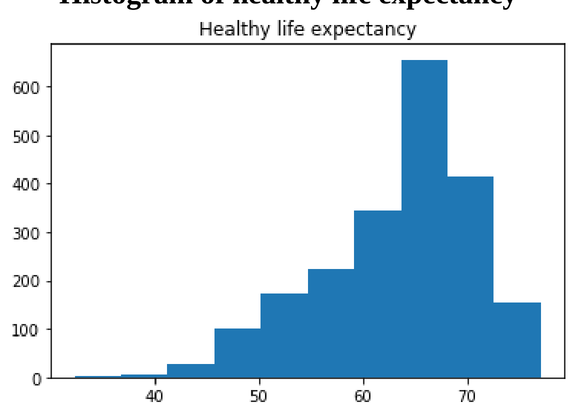
Appendix B.3

Histogram of social support



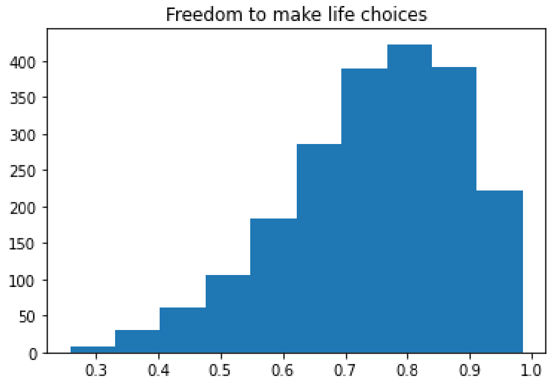
Appendix B.4

Histogram of healthy life expectancy



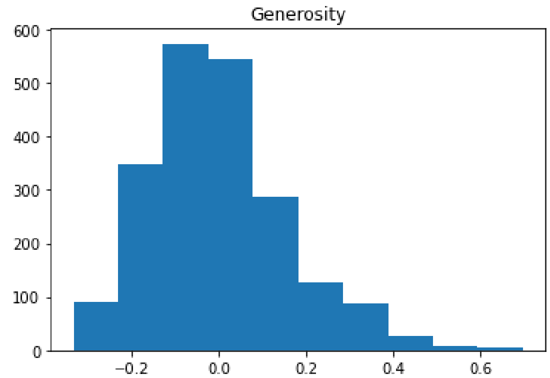
Appendix B.5

Histogram of freedom to make life choices



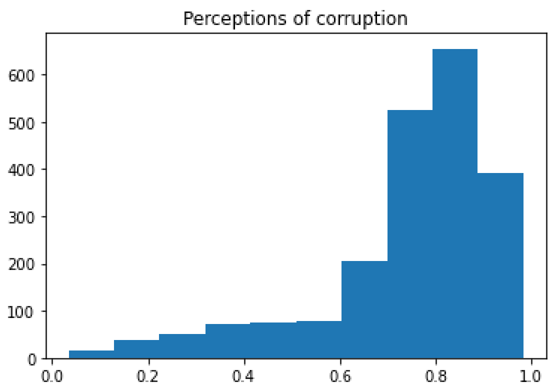
Appendix B.6

Histogram of generosity

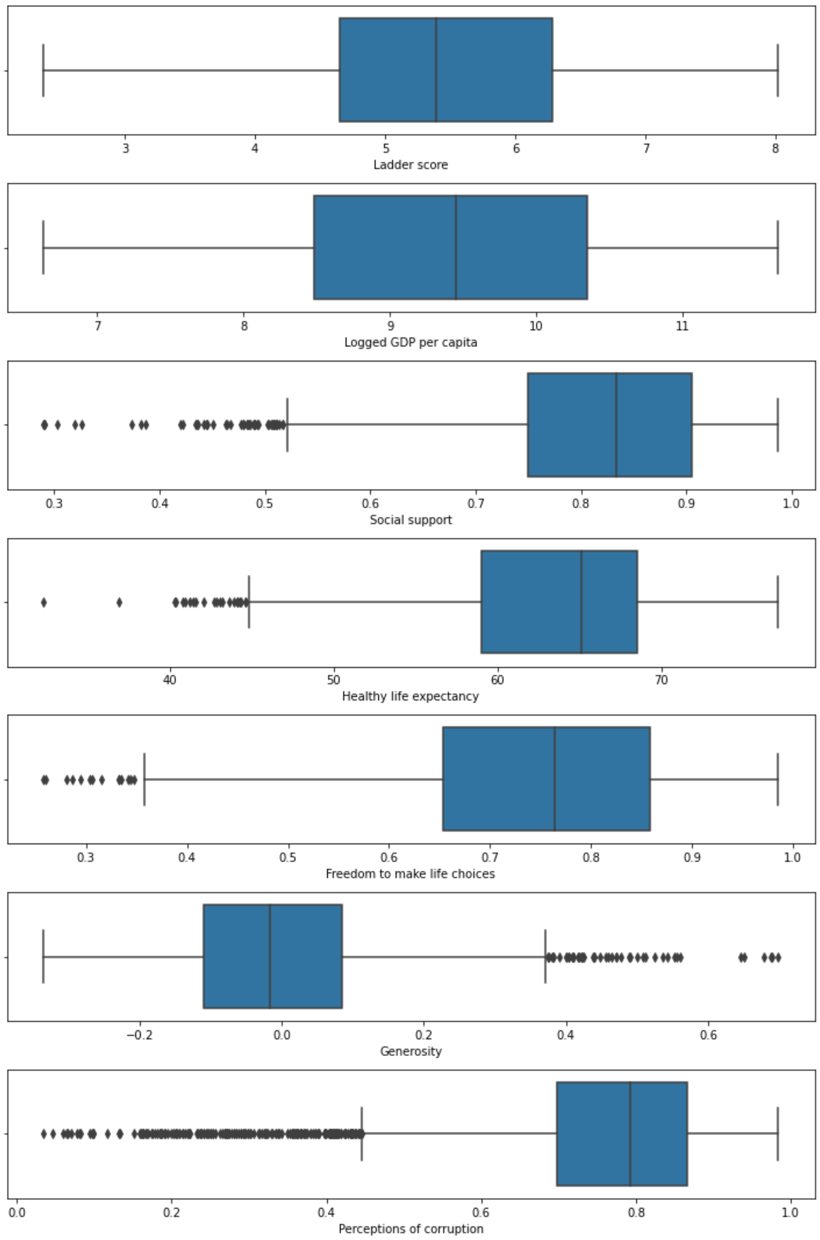


Appendix B.7

Histogram of perceptions of corruption

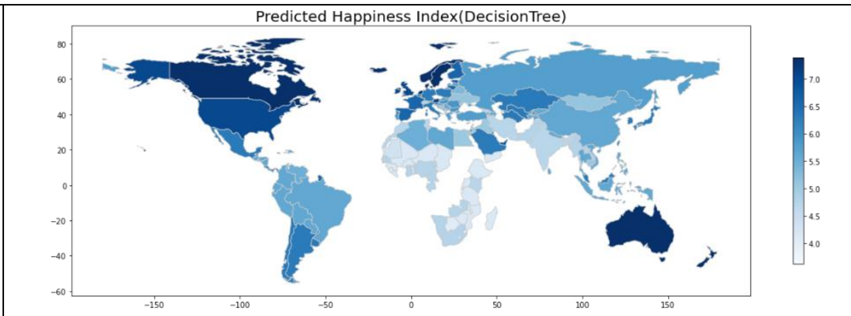


Appendix C Boxplots



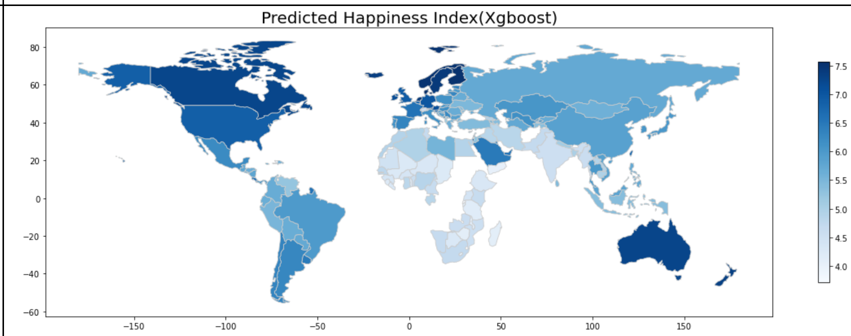
Appendix D.1

World Map of Predicted Happiness Index from Decision Tree



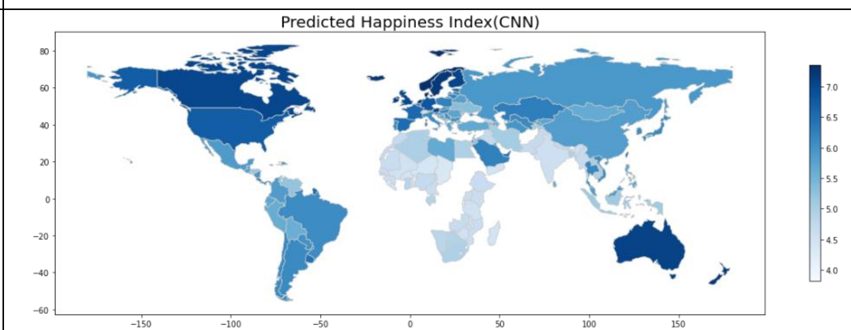
Appendix D.2

World Map of Predicted Happiness Index from XGBoost



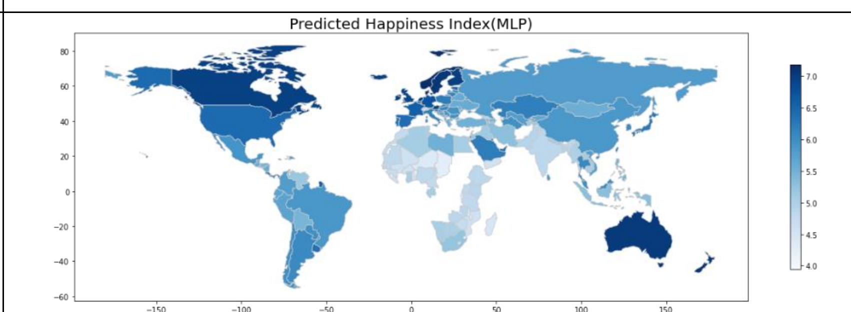
Appendix D.3

World Map of Predicted Happiness Index from CNN



Appendix D.4

World Map of Predicted Happiness Index from MLP



9. References

Anshul.S (2021)An Introduction to Random Forest Algorithm for beginners, from https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners

Baumeister, R. F., & Leary, M. R. (2017). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. Interpersonal Development, 57–89. https://doi.org/10.4324/9781351153683-3

Britannica, T. Editors of Encyclopedia (2021, December 17). Happiness. Encyclopedia

Britannica. Retrieved May 6, 2022, from https://www.britannica.com/topic/happiness

Brownlee J. (2021, March 12) XGBoost for Regression. From https://machinelearningmastery.com/xgboost-for-regression/

Brulé, G., & Veenhoven, R. (2014). Freedom and happiness in nations: Why the Finns are happier than the French. Psychology of Well-Being, 4(1). https://doi.org/10.1186/s13612-014-0017-4

Diener, E., & Seligman, M. E. P. (2002). Very happy people. Psychological Science, 13(1), 81–84. https://doi.org/10.1111/1467-9280.00415

Dinkar, P. (2021, June 10). World is happy or is it? EDA + ml. Kaggle. Retrieved April 14, 2022, from https://www.kaggle.com/code/sonukiller99/world-is-happy-or-is-it-eda-ml

Easterlin, R. A. (2012). Happiness and economic growth: The evidence. IZA.

Jason.B(2018) How to Develop Convolutional Neural Network Models for Time Series Forecasting, from https://machinelearningmastery.com/how-to-develop-convolutional-neural-network-models-for-time-series-forecasting/

Mahapatra, S. (2018, May 22) Why Deep Learning over Traditional Machine Learning? from https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

Philip. Yam & Kaiser. Fan (2022) Lecture note of STAT 4012 Statistical Principles of Deep Learning with Business Applications

Richins, M. L., & Rudmin , F. W. (1984). Materialism and Economic Psychology. Journal of Economic Psychology, 1(5), 217–231. https://doi.org/10.1016/0167-4870(84)90035-7

Thompson, D. (2019, July 15). The happiness dividend: Longer, healthier lives. WebMD. Retrieved May 6, 2022, from https://www.webmd.com/healthy-aging/news/20190715/the-happiness-dividend-longer-healthier-lives