Fall 2020 CIE 417 Saturday, December 26th, 2020

**Project Phase 3: Final Report**

Alaa Alajmy (ID # 201700095)
Moemen Khaled (ID # 201700873)

# Understanding and Predicting Happiness: A Machine Learning Approach

## Contents:

# I. Introduction

As said by Aristotle, “happiness is the meaning and purpose of life, the whole aim and end of human existence”. Yet, eras of human existence notwithstanding, the nature of “happiness” still eludes appropriate understanding. While Aristotle says happiness is gained by gaining health, wealth, and friends, Socrates claims happiness is gained not by gaining more, but by wanting less.

Paul Lazarsfield, a mathematician and professor of sociology, wrote you could theorize education is positively related to stress in tough conditions since less educated people are more likely better adapted to tough conditions and simultaneously, you could theorize education is negatively related to stress in tough conditions since more educated people are more likely better informed about coping with stress.

It is this confounding maze that leads us to forgo intuition and common sense and depend on data and results to better understand what factors affect the happiness of the average human.

## a. Problem Definition and Motivation

In this project, we will use machine learning methods to quantitatively define the main factors affecting humans’ subjective sense of happiness and thus develop a method to predict happiness. Besides sating thousands of years of human curiosity, the results of this project could aid in the sustainable development of the lives of communities and nations. We will specifically seek to understand the causal effect of demographic, economic, political, and social influences on humans’ sense of happiness.

## b. Previous Work

To the best of our knowledge, no work has been previously done to try to predict human happiness using demographic, economic, political, and social influences. Similar work has been done to analyze the factors influencing happiness using machine learning in [1] and by the United Nation’s World Happiness Report team. Additionally, several reports have analyzed the correlations between happiness and several environmental influences using machine learning and other techniques [2].

# II. Methodology

## a. The Dataset

We will use the UN's World Happiness Report (WHR) datasets created over the years 2005 till 2020. The WHR data is created by taking the average per nation of data collected through the Gallup World Poll survey, which correlates demographic, economic, political, and social factors with real-world outcomes.

The Gallup World Poll (GWP) is conducted yearly in at least 153 countries with a random sample of at least 1000 respondents in each country.

Our compiled dataset could be found here: https://drive.google.com/file/d/1Bc05x4ciSbjeqhE7w0qJRihlssUSZL89/view?usp=sharing.

Each entry in the dataset is labeled by a country and a year. These two features are merely labels and hence are of no importance to our algorithm and will be ignored. The remaining features are:

1. **Subjective Happiness (named Ladder in the dataset):** this is a measure of average subjective happiness on a scale from 0 to 10 where 0 is least happy and 10 is most happy. This is our target feature.
2. **Log GDP per capita (LogGDP)**: this is the log of each entry's gross domestic product divided by its population. The GDPs are retrieved from data created by the World Bank.
3. **Social support (SocialSupport)**: this is the average of responses to the question "Do you have someone to count on in times of trouble?". In the GWP, this question was answered by either 0 or 1, thus the averaged values in our dataset per country are in the range 0 to 1.
4. **Healthy life expectancy at birth (HealthyLifeExpectency)**: this is each entry's average life expectancy in good health. This data was extracted from reports by the World Health Organization.
5. **Freedom to make life choices (Freedom)**: this is the average of responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?". As in the social support feature, the values are in the range 0 to 1.
6. **Generosity (Generosity)**: this is the residual (the vertical distance between a data point and the regression line) of regressing national average of responses to "Have you donated money to a charity in the past month?" on GDP per capita. This feature thus ranges from -1 to 1. (To clarify this feature's significance, a residual is the error that isn't explained by the regression line, thus generosity is the donations or their lack that do not fit into the expected donations vs. GDP regression line.)
7. **Perception of corruption (Corruption)**: this is the average of responses to the question "Is corruption widespread throughout the government or not?" and the question "Is corruption widespread within businesses or not?" . The two questions are averaged for each GWP respondent to produce a value of either 0, 0.5, or 1, thus the range of this feature is 0 to 1.

8. **Positive affect (PositiveAffect)**: this feature represents the average of whether respondents experienced laughter or enjoyment a lot during the past day. The range is again from 0 to 1.
9. **Negative affect (NegativeAffect)**: this feature represents the average of whether respondents experienced worry, sadness, or anger a lot during the past day. The range is from 0 to 1.

We most point out that, as explained above, of this data is collected through surveys and is thus mainly subjective. That is, when we refer to happiness or corruption below, we mean the averaged person's perception of his or her own happiness or the averaged person's perception of corruption in the government or in businesses

## b. Ethical Consideration

In this project, we use the UN's World Happiness Report (WHR) datasets created over the years 2005 till 2020. This data is created by taking the national average of the Gallup World Poll's results. The Gallup World Poll (GWP) is conducted yearly in at least 153 countries with a random sample of at least 1000 respondents in each country. The survey is conducted over the phone in countries where most of the general population are phone users and face-to-face in others. All interviewees receive intensive training to ask questions in a non-leading manner and the interview is given in the languages most used in each country. The Gallup World poll is thus sufficiently representative of 95% of the world's adult population

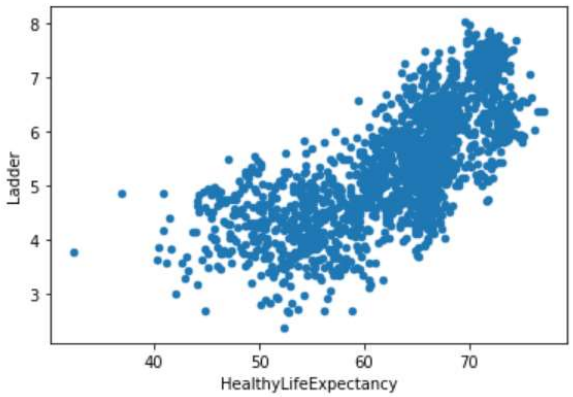
The WHR's final features are all either measures of subjective experiences (e.g. positive affect and social support) or objectively recorded facts (e.g. GDP and healthy life expectancy at birth).

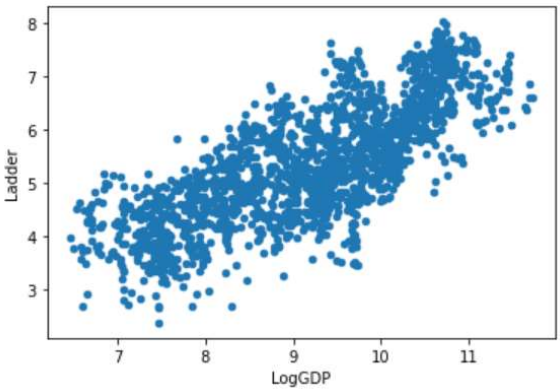
## c. Data Exploration

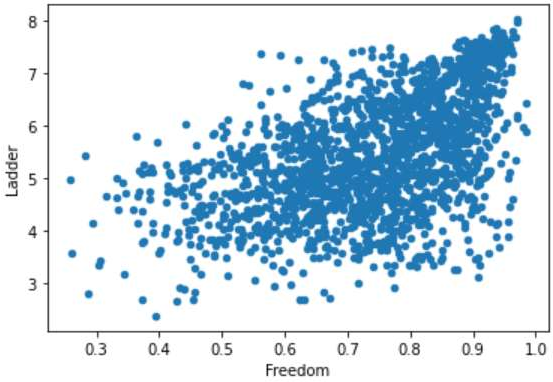
Before applying machine learning algorithms, we initially explore the correlations explicit in our data. We ask questions on how different environmental factors influence happiness, and we ask more diverse questions on how different environmental factors influence one another. Below are some of our results:

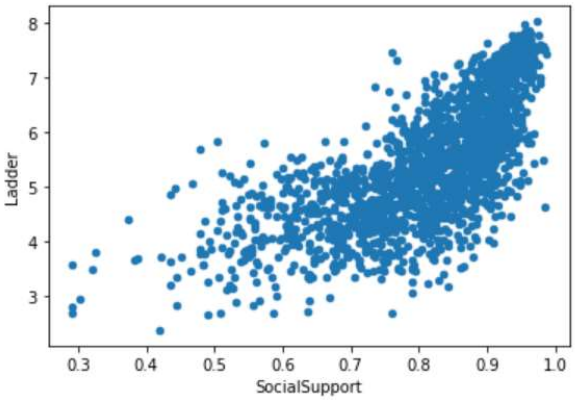
### i. The Influence of Environmental Factors on the Happiness Ladder

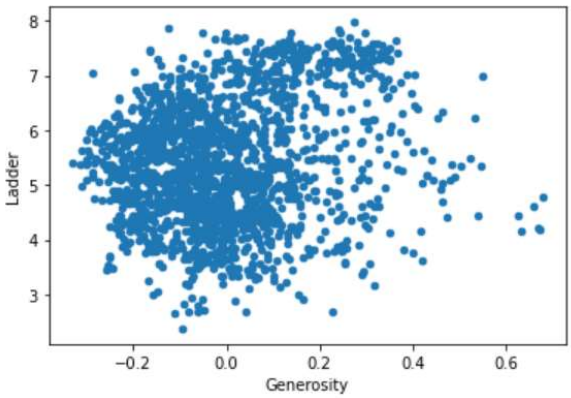
As seen in the plots below, we found GDP, healthy life expectancy, and social support are strongly positively correlated with happiness. Freedom also shows some positive correlation with happiness. Generosity shows almost no correlation and only in countries with low levels of corruption perception, happiness and corruption perception are negatively correlated. Expectedly positive affect is positively correlated with happiness and negative affect is negatively correlated, yet these two plots show weaker correlations than might be expected.

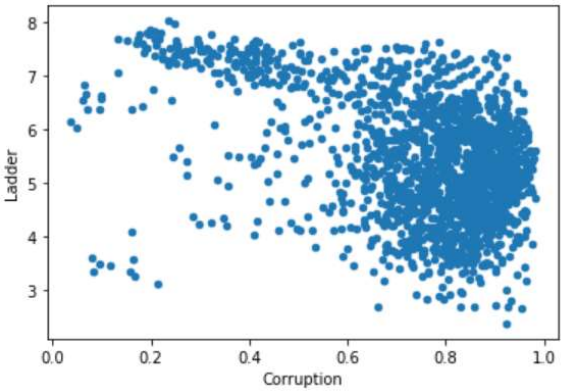

Ladder
HealthyLifeExpectancy

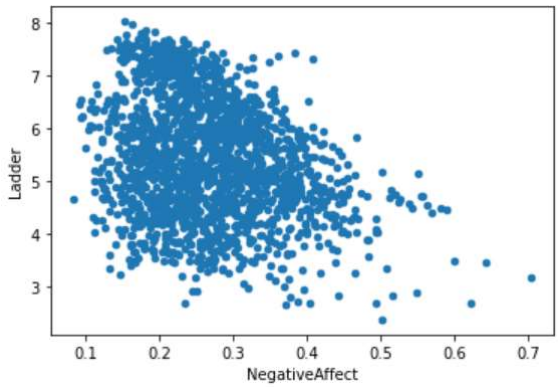

Ladder
LogGDP

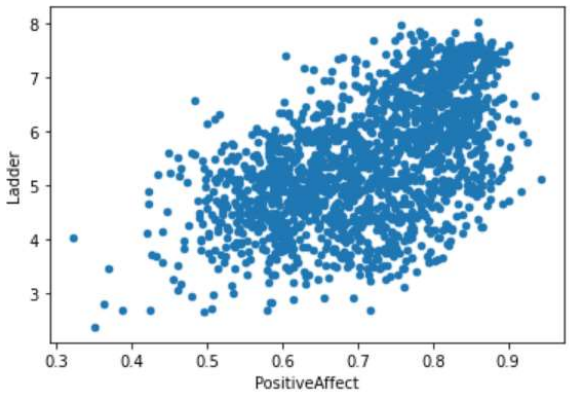

Ladder
Freedom


Ladder
SocialSupport


Ladder
Generosity


Ladder
Corruption

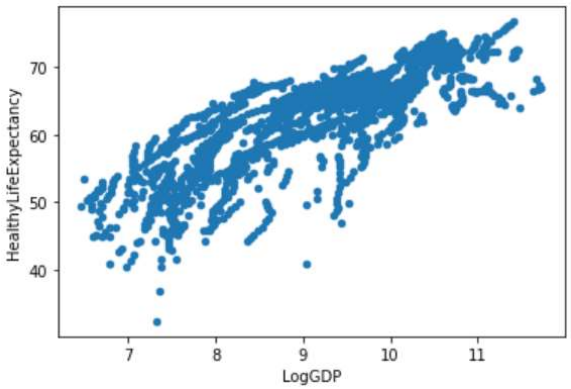


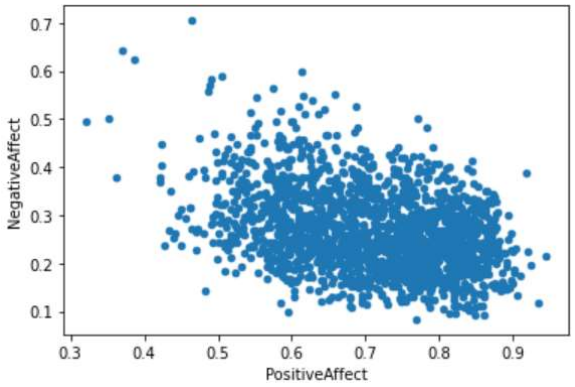



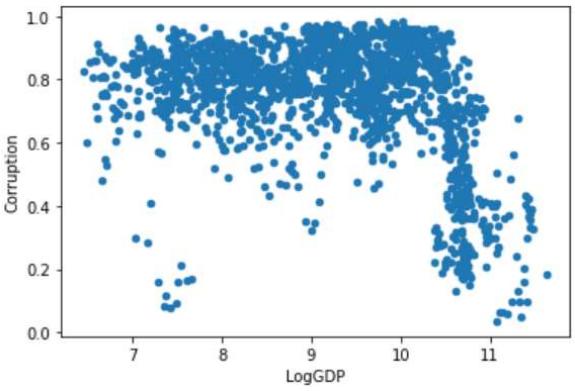

ii. The Influence of Environmental factors on One Another

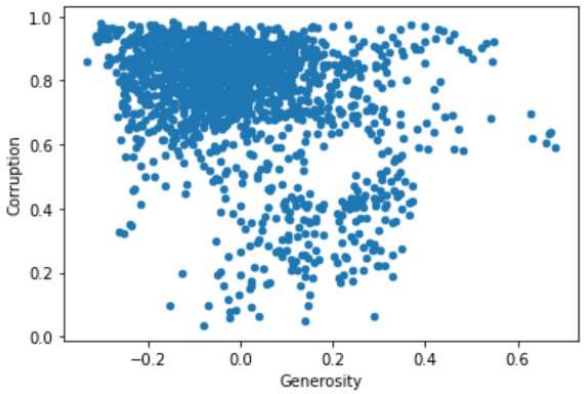
We ound GDP to be positively correlated with healthy life expectancy and positive and negative affect to be weakly negatively correlated. We found no correlation between corruption and GDP or corruption and generosity. We found an odd slight negative correlation between GDP and generosity.

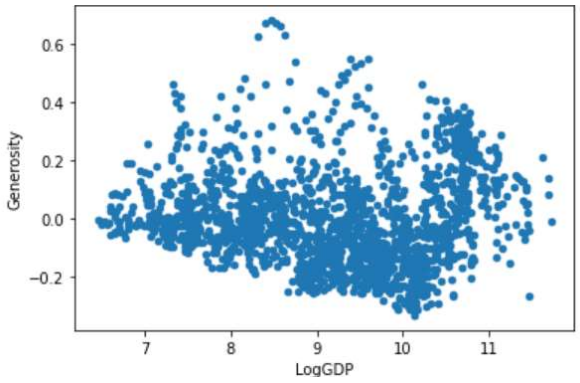
At the end, we found that most coutries have a lower than average GDP and a lower than average healthy life expectancy.

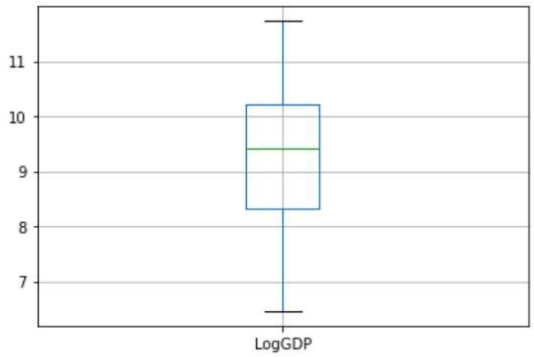


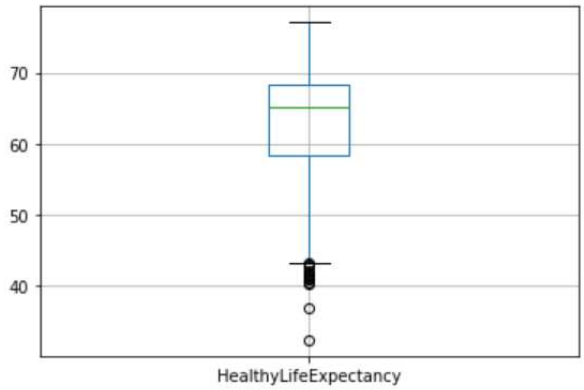






## d. Machine Learning

This is a regression problem so the first algorithm that comes to mind to tackle it is Linear Regression. We will apply linear regression using ridge and lasso regularization and tune their hyperparameters with k-fold cross validation. We will also try to apply linear regression after using PCA and RFE to filter out our features. We will also apply neural network models of depths 3, 4, and 5 along with KNN regression and decision tree regression.

Before applying any regression algorithm, we initially drop all entries with missing social support, negative affect, or positive affect values. We fill in all remaining missing values using KNN imputation with k = 40, approximately the square root of our data count. At the end, we are left with a dataset of 1823 rows and 9 features, one of which is our target.

We generally aim at an R2 score of 0.9. After determining the best algorithm to use, our project will quantitatively display the demographic, economic, political, and social factors that affect human happiness and will be able to receive the attributes belonging to some country and predict its average subjective happiness. Below is a summary of our steps and conclusions. A detailed documentation of our code and results could be found in the attached notebook.

i. Linear Regression

We apply on our data linear regression, initially using no regularization, then using lasso and ridge regularization. We tune our regularization factor using Sklearn's default leave one out method. We also apply Principal Component Analysis (PCA) linear regression and use Recursive Feature Elimination to fit a model with only 4 features. We try both normalizing and standardizing our data and find standardization to yield significantly superior results.

Our best-performing linear regression model was standard un-penalized linear regression using standardized data. This model gave an R2 score of 0.761 and gave the highest weights to **GDP, social support, healthy life expectancy, and positive affect** in that order. This models had an intercept of 5.42756807169906 and coefficients of:

GDP: 0.3956963987589417
Social Support: 0.25185293335633524
Healthy Life Expectancy: 0.20995529824571735
Freedom: 0.07077083209899611
Generosity: 0.06910445609437851
Corruption: -0.11696266386826576
Positive Affect: 0.1936845582189926
Negative Affect: 0.007735313462046325

Our ridge regression model, also using standardized data, gave similar results. All lasso regression models performed badly. The best performing model gave a score of 0.54 using non-standardized, non-normalized data. This terrible performance may be because lasso regression is generally well-performing in features having high multicollinearity only, and apparently our dataset does not fall under that umbrella.

Our best-performing PCA regression model, tuned using a validation set, used 7 principal components and gave the highest weights to GDP, social support, and healthy life expectancy.

Using RFE, we were able to create a model with only the 4 best-performing features and an R2 = 0.752. This model used (in order of highest-to-lowest regression weights): **GDP, positive affect, social support, healthy life expectancy**. The intercept here was 5.42756807169906 and the coefficients were:

GDP: 0.42827193
Social Support: 0.2310698
Healthy Life Expectancy: 0.22872935
Positive Affect: 0.29027058

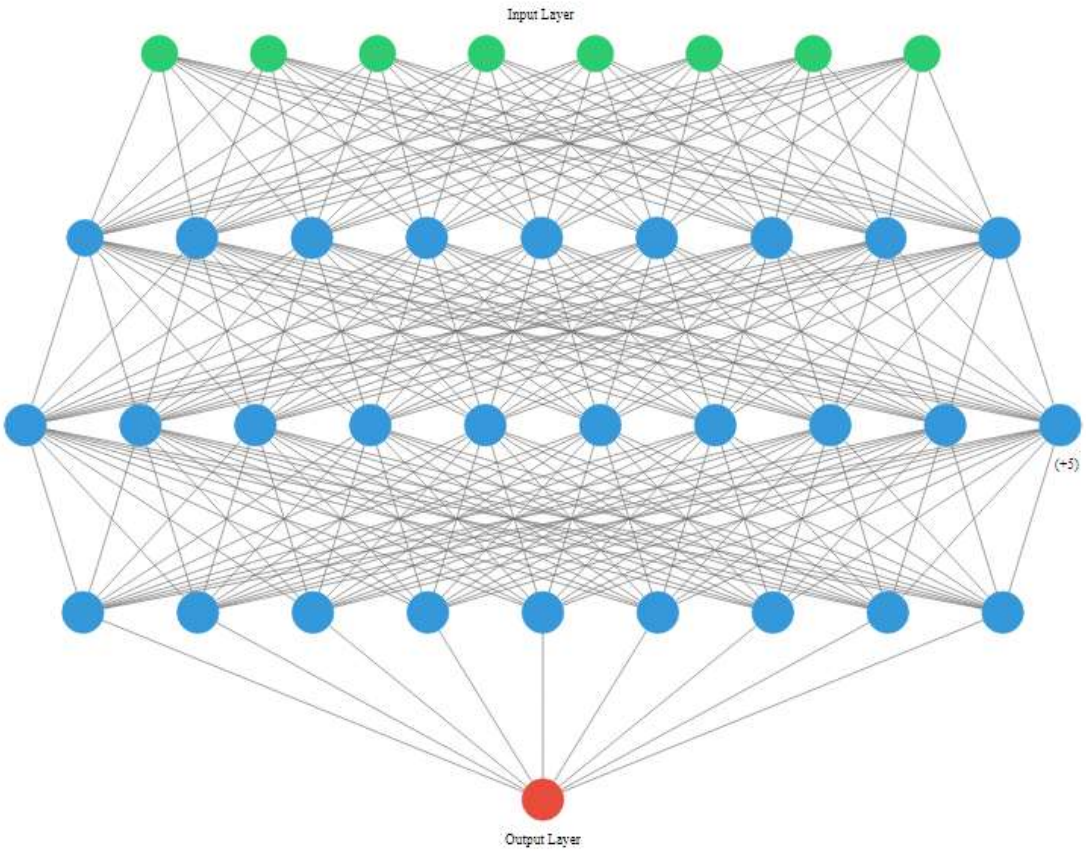
ii. Neural Networks

We create several single-neuron models, simple neural networks with three layers, more complex networks with four layers, and complex network with 5 layers. We use min-max scaled data and use R2 scores evaluate our models. For each of the neural networks we train, we fit the model several times then average the R2 of all trials to overcome the randomness in the fitting algorithm.

To reach the best optimization of the network weights, we try with each model different optimizers and pick the one with the highest reliability. Our used optimizers are tensorflow's RMSprop, sgd, and adam.

We define the reliability of the optimizer as the average R2 score of the test data produced from training multiple models. Once the most reliable optimizer is found, we pick the model with the highest R2 that it produced.

Our best neural network model is the four-layer simple network with two Exponential and one Softmax activation functions, trained using the SGD optimizer. This model gives an R2 = 0.7870 and looks like this:




For single neuron networks, the best model is a neuron with a ReLu activation function and an R2 = 0.7052, trained using the SGD optimizer. For three-layer simple networks, the best model has one Exponential activation function, trained using the SGD optimizer, and an R2 = 0.7808.

iii. KNN Regression

Using k-fold cross-validation with k in the range 3 to 5, we fit a KNN regression model that uses 3 neighbors and returns an R2 score of 0.8482. The model weights points by the inverse of their distance and uses the Manhattan distances metric.

iv. Decision Trees Regression

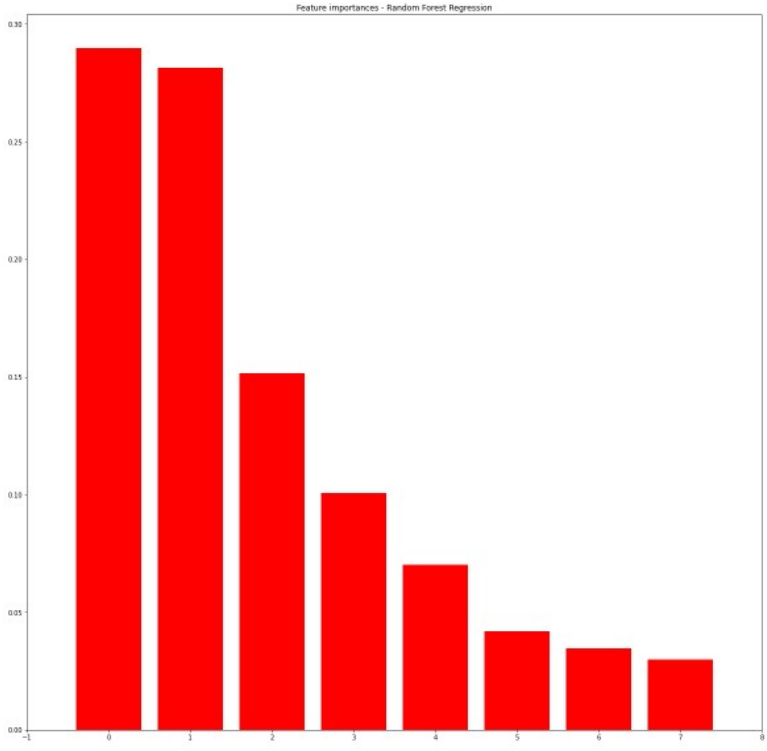
We train a simple regression tree, a random regression forest, and an Adaboost regressor.

Our best decision tree model has an R2 = 0.8674 and is a random regression forest trained with the following parameter:

Criterion: Mean Squared Error (MSE)
Max Features: log2 of the total number of features
Number of Estimators: 110
Warm Start (Using previous tree solutions as an initial guide): True

This model gives each feature a relative importance of:

1. Healthy Life Expectancy: 0.28963400880212714
2. Log GDP: 0.28142791753335933
3. Social Support: 0.151690113514482
4. Positive Affect: 0.10074182048740042
5. Freedom: 0.07028997879232103
6. Corruption: 0.0418316314903682
7. Generosity: 0.034591084536810465
8. Negative Affect: 0.02979344484313155



As apparent in the histogram above, the random forest regressor gave the highest value to **healthy life expectancy, log GDP, social support, and positive affect.**

Second best was the Adaboost model with R2 = 0.7987. This model had:

Learning Rate: 0.2
Loss Function: exponential
Number of Estimators: 120

The Adaboost model gave the highest importance to **healthy life expectancy, log GDP, positive affect, and social support,** in this order.

Using all seven features, our simple tree yielded an R2 score of 0.7787 and a maximum depth of 6. The simple tree gave the highest importance to **log GDP, healthy life expectancy, positive affect, and social support**, in this order. Using only the four most important features, the simple tree returns an R2 = 0.7747 with a maximum depth of 7.

## III. Results

Although the order in which they are valued tends to vary among our regressors, **healthy life expectancy at birth, GDP social support, and positive affect** are, according to all our well-performing models, the best predictors of subjective human happiness. We found that these four features vary positively with the happiness ladder.

Our best-fitting model, the random regression forest, gives the highest values to healthy life expectancy at birth and log GDP and almost half these values to social support and positive affect. Meanwhile, our best performing linear regression models gave the highest values to log GDP.

Overall, our best performing model is the random forest regressor which uses 110 estimators and yields an R2 score of 0.8674.

## IV. Discussion

While our work yields quite promising results, human happiness is most definitely a goal more intricate than 8 measures of environmental influences.

For machine learning engineers, many exciting paths are open for investigation. For instance, while the dataset we used was the largest, most comprehensive, and least biased dataset available to us, a future model could make use of a more expansive dataset with additional features including a country's political condition, the status of its minorities, and the actual economic status of its citizens. We did find some, yet not all, of these features in other surveys, most prominent of which was the World Values Survey (WVS). Unfortunately, the WVS was much too small and had

alarmingly many missing values, making too hard the possibility of incorporating it into our WHR dataset.

In the words of Gautama Buddha, “there is no path to happiness: happiness is the path”.

## V. References


[1] L. Millard, “Data Mining and Analysis of Global Happiness: A Machine Learning Approach”, 2011. [Online]. Available: http://www.datamining.org.uk/MSC_THESIS_FINAL_VERSION.pdf. [Accessed: 26-Dec-2020].

[2] L. Faik, “Understanding Happiness Dynamics with Machine Learning (Part 2),” *Medium*, 23-Aug-2020. [Online]. Available: https://towardsdatascience.com/understanding-happiness-dynamics-with-machine-learning-part-2-4df36e52486. [Accessed: 20-Nov-2020].

[3] “Unbiased news powered by public opinion research and analysis of human behavior around the world,” *Gallup.com*, 19-Nov-2020. [Online]. Available: https://news.gallup.com/home.aspx. [Accessed: 20-Nov-2020].

[4] https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

[5] https://jmlb.github.io/ml/2017/03/20/CoeffDetermination_CustomMetric4Keras/

[6] https://keras.io/api/layers/activations/

[7] https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6

[8] https://www.tensorflow.org/api_docs/python/tf/keras/optimizers

[9] https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/

[10] https://stackabuse.com/decision-trees-in-python-with-scikit-learn/

[11] https://towardsdatascience.com/visualizing-artificial-neural-networks-anns-with-just-one-line-of-code-b4233607209e