Title: World Happiness Report

Name: Ayushi Mishra

We all aspire to find things that really makes us happy, yet the key factors behind it are still unclear. I was curious to know what features would lead to global happiness but formula for calculating it is confidential, I decided to use Machine Learning techniques unravel the formula of global happiness. For my analysis I first gathered "The World happiness Report" dataset from Kaggle, https://www.kaggle.com/unsdsn/world-happiness, which measures global happiness based on Gallup World Poll (GWP). I also used Global Health Observatory (GHO) dataset "Life Expectancy (WHO)" https://www.kaggle.com/kumarajarshi/life-expectancy-who from Kaggle, which provided me contributing factors towards happiness in order to gain a broader picture for my analysis. This report is useful for Global leaders to effectively track metrics behind it and implement ways to lead progress of their nation. For this report, I merged the two datasets by first filtering out only year 2015 from both and cleaned up the data, as population, BMI, Total expenditure and alcohol consisted of missing values. The final dataset that I used for my analysis consists of 135 rows and 19 columns.

We will start by looking at **Figure 1**, which shows how the average values of the four selected features for my model, contributing towards Happiness score for each region, varies. For this plot, I created a new dataframe, which was a subset of the original one, consisting of the aggregate mean of these four features, which were grouped by region. I also added a new column for Happiness score by converting it in the same scale (i.e. out of 1) as other features, to see how much does the height for Happiness score differs for each region in presence of these contributing factors. Using region on my x-axis was a useful to analyze the Happiness Score as there seems to be significant difference across developed, emerging and developing regions. Looking at the plot, we can infer that average values for Income composition and Life Expectancy tends to have a high average value across all the regions as compared to freedom and trust. If we look at the bars representing Happiness score, it shows a general trend which correlates with the average values of Income composition and Life Expectancy for that region. This can be observed by looking at regions like Eastern Asia and Central and Eastern Europe almost equal score, that even though the average values for trust and freedom seems to be quite low but still the average Happiness score is high. Also, on observing the bars for regions like North America, Australia and New Zealand and Western Europe, the Happiness score tends be highest, due to higher income composition and life expectancy (close to 1), implying that highly developed regions generally have higher happiness score. While, Sub-Saharan Africa has lowest happiness score due to lowest average values of all 4 contributing features, implying least developed regions have a lower happiness score.

Next for **Figure 2**, we used Principal Component Analysis to find smallest number of representative features to describe dimensionality of the data. The yellow line shows that without scaling the data we need atleast four features to achieve explained variance of 98%. On applying StandardScaler on our data, we can see that the blue line follows a similar pattern as that of without scaled data, mainly because our feature values lie in the same range (0 - 1). Hence, using PCA we can reduce our dataset from 8 to 4 features to explain around 83% of variance.

To further continue my analysis, I decided to find which of the features tend to have larger contribution in determining the score. But to achieve this, I used sklearn (1) LinearRegression model by first performing 75%/25% train and test split on our data and I observed that my model achieved a score of 95.4%. Then I also created sklearn (2) pipeline with Standard Scaler for preprocessing my data along with LinearRegression, which gave me the exact score as previous (95.4%). This could be possible because all features of our data seem to be on the same scale. In order to improve my model, I performed a pipeline on my dataset by using (3) PolynomialFeatures with degree 2, and I observed that my model achieved 89.7% score, which is lesser than before. I also used cross_val_score as our metrics to check the accuracy of (1) model and (3). I observed that (1) seems to give a mean score of 94% and variance close to 0 and hence shows that our model is not overfitting. But the 3) model gives a mean score of -2.128 and variance equal to 3.64, which definitely indicates that the model is overfitting. Hence, to understand coefficient weights for each feature, shown in **Figure 3**, I will be using my Linear Regression model. The plot shows that freedom tends to have the highest weight of 1.6, and hence it is positively correlated with Happiness score. GDP per capita and generosity are other influential factors behind the model. In conclusion, a possible formula from the top 6 most influential factors for calculating happiness score would be: generosity * 1.2 + life expectancy * 0.65 + freedom * 1.6 + GDP per capita * 1.3 + income composition * 0.8 + dystopia residual * 0.9. In conclusion, this model gives us an idea of the policies and incentives must be implemented by government. This includes providing enough freedom to make life choices and healthy living conditions to citizens. Also improving trade surplus to increase GDP.

Figure 1: Average of Features that contributes to Happiness for each Region

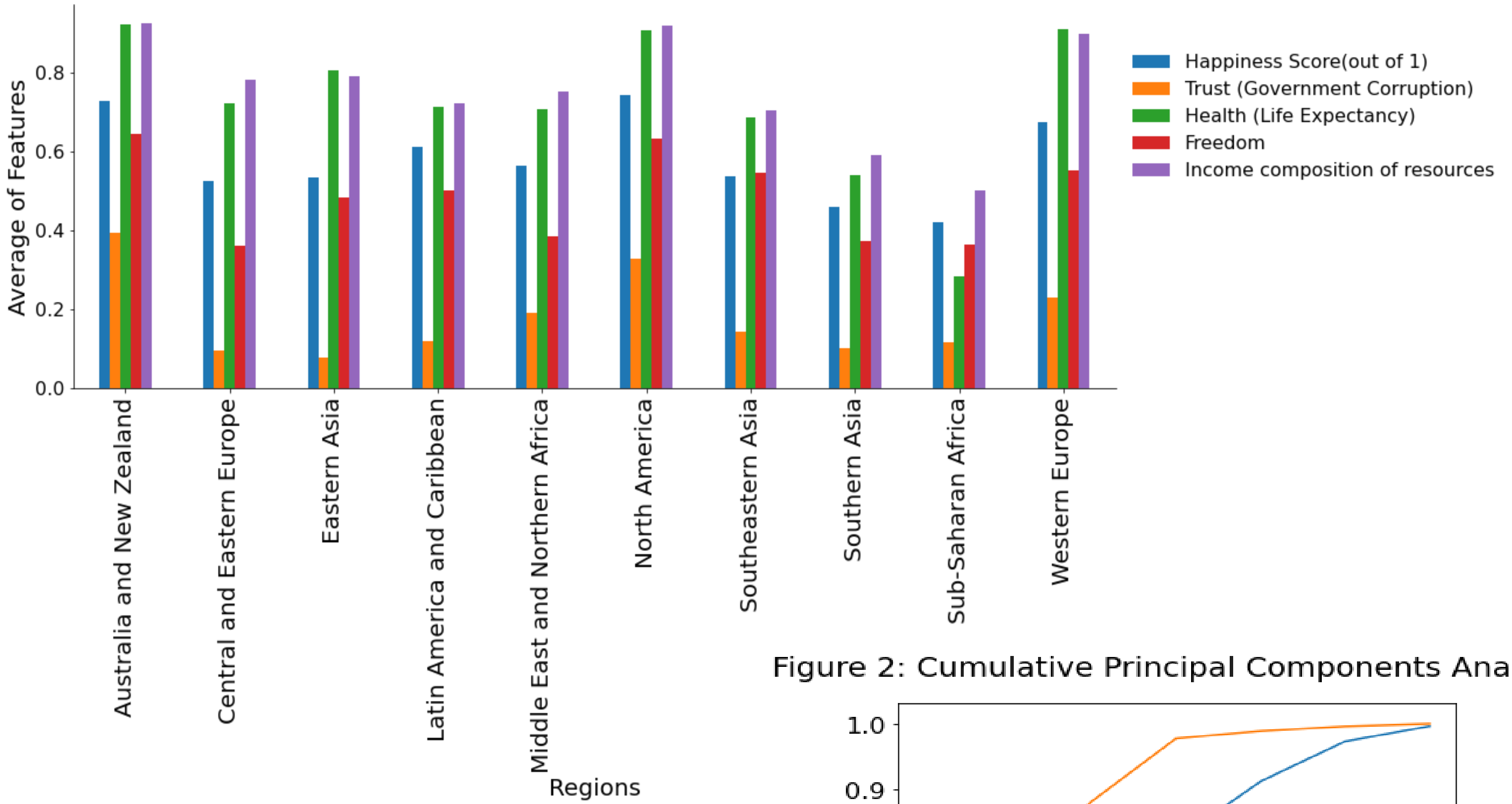


Figure 2: Cumulative Principal Components Analysis

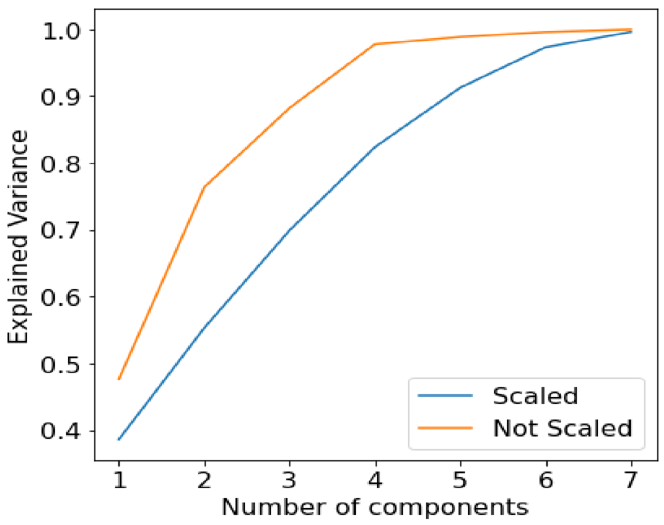
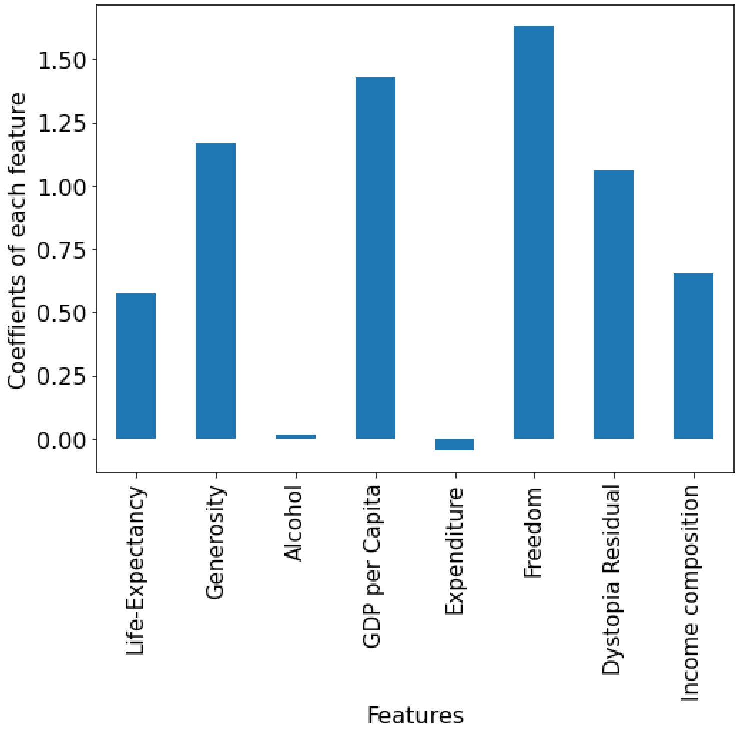


Figure 3: 8 Most Significant Features that explains the Linear Regression model


Coeffients of each feature
Life-Expectancy
Generosity
Alcohol
GDP per Capita
Expenditure
Freedom
Dystopia Residual
Income composition
Features