Aditya Chopra, Dhruv Gandhi, David Ueda, Benson Yu

Professor Kevin Gold

DS110 Introduction to Data Science with Python

12 December 2022

**World Happiness Report**

**Dataset:** https://www.kaggle.com/datasets/unsdsn/world-happiness?resource=download

**Introduction**

Whilst trying to find our dataset we came across the fact that our group members have backgrounds from different parts of the world. This encouraged us to ponder about differences in the world which can be measured through data. We came across the ‘World Happiness Report’ Dataset and wondered which parameters would affect happiness the most and if these measurements could draw predictions accurately. The ‘World Happiness Report’ dataset gave us values of happiness score and rank, given its characteristics of the country such as GDP and health of the country. Furthermore, the data is comprehensive in terms of how they conclude a happiness score, allowing us to vary more parameters and apply more approaches to analyze the data. Using this dataset, our goal was to construct different models to analyze the happiness scores of different countries and experiment to what extent happiness can be predicted using machine learning in Python. The two approaches used to analyze the dataset are the K Nearest-Neighbors Classification model and the Linear Regression model.

**Methodology**

KNN Classification Model

The World Happiness Score Report included datasets from the years 2015 - 2019. All of the datasets provide us with the Country, Happiness Rank, Happiness Score, and many other characteristics about each country. We planned to use all the datasets by concatenating all the data from 2015-2018 and using a KNN regression model with the data between 2015-2018 to predict the happiness score for 2019 and see how accurate we would get. However, we realized that each dataset has the same countries and that it might not be fair to make predictions if the same country, with the same features, appears with basically the same happiness score across multiple years. Furthermore, our other model was also going to be a regression model with basically the same idea. Instead of making two regression models, we decided on making a KNN classification model. Since the 2016 dataset had a Region characteristic that the others did not, we decided to use the 2016 dataset for that reason. It would allow us to be able to classify each data point by its Region given its other characteristics. It would also fix the problem we ran into earlier by using only a year instead of multiple years combined.

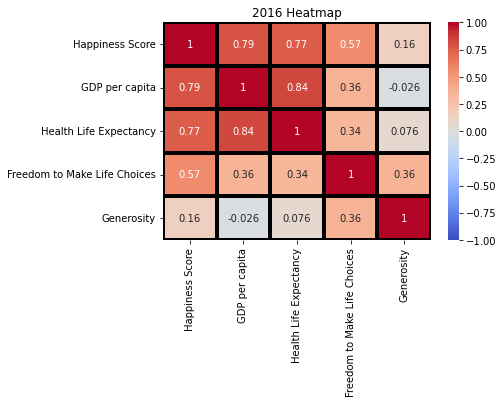
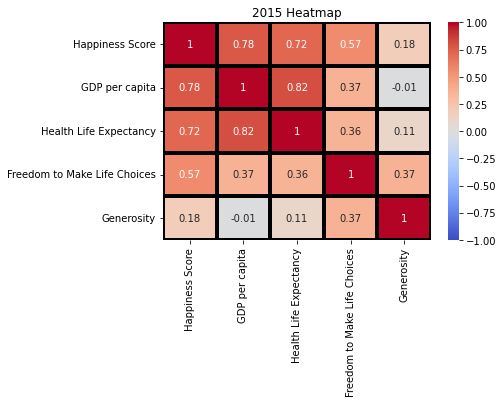
We began by importing our dataset and cleaning the data for our model. We wanted to have certain characteristics about each country for example, Region, Happiness Score, and GDP. First, we dropped columns that were not needed and then we made Region our target variable. Then we checked for any imbalance data in our variables to see if we had to add or remove any data points, which we did. We decided to upsample our data because it is better to have more data points (Thorson 2021). First, we looked at how many data points were in each category. Then we defined each target category as a class and used the sample feature to add data points that are randomly sampled from each class distribution until each class had the same number of data points.

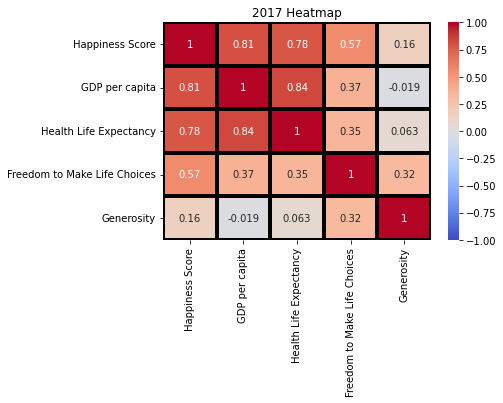
Finally, we were able to start training our model. We used the MinMaxScaler package from Sklearn to scale our input variables from zero to one so it does not throw off the model. An important part of the KNN classifier is the number of neighbors; we needed to know the optimal K value to use. This value was calculated by square rooting the number of samples, which was 157, to get 13 (Band 2020). Finally, we set the number of neighbors equal to 13 and ran our model.

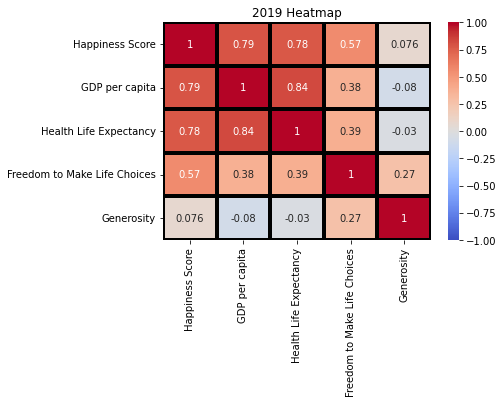
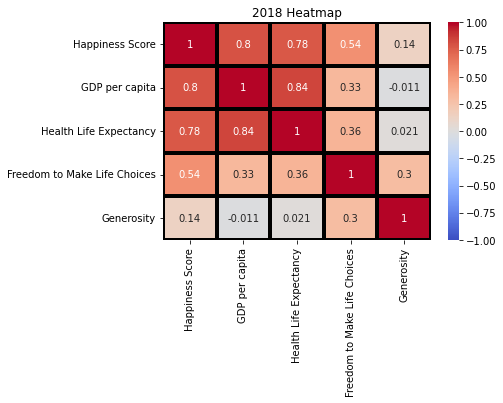
Linear Regression Model

Creating a linear regression model for the world happiness score report required a lot of data cleaning. The reason for this being, at first we thought to concatenate all the data from 2015-2019 to create a basis to run linear regression. However, we realized that doing this could compromise the model as the scores for many countries were extremely similar from 2015-2019. This would mean that while training our data to find the predicted happiness score, there would be a high likelihood for the predictions to show almost exactly the same data. At the same time, however, we were still curious to see if there were any underlying differences between the years. from year to year. This is why we decided to train and test a linear regression model for each year separately.

Before we created the linear regression model, we decided to test the correlations between the parameters we and the happiness score for each year. We did this by creating heatmaps for each year’s data. Seaborn’s heatmaps are good visual tools to assess the relationships between different variables. In our heatmaps below, the correlation between the parameters and the happiness score are determined through a ‘r’ score. This helped us to understand which parameters had the highest effect on happiness score and which ones had the lowest affect.





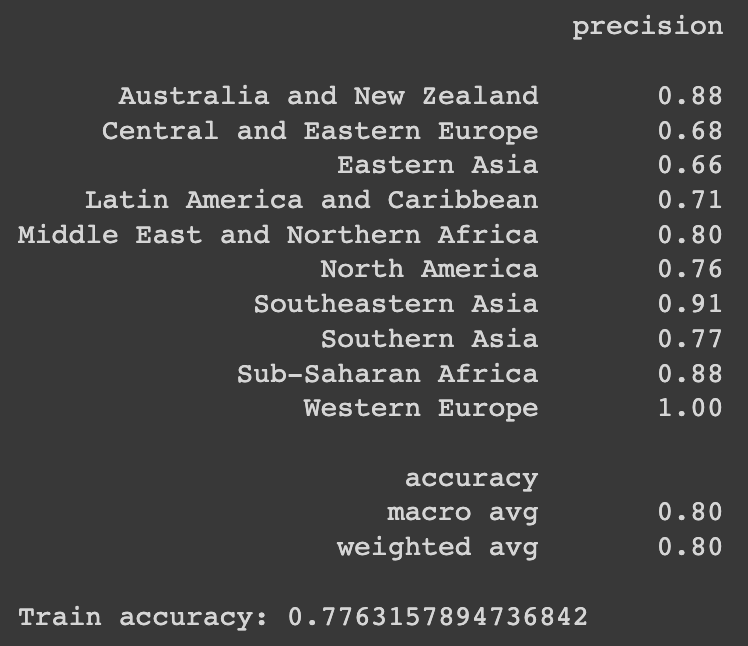


The heatmaps above show us that GDP per capita and Health Life Expectancy had the highest correlation to happiness score for every year. Ergo, these two parameters have the highest effect on happiness score and its predictions. We can also see that Freedom to Make Life Choices and Generosity had the lowest correlation where Generosity was significantly lower with almost no correlation to happiness score.

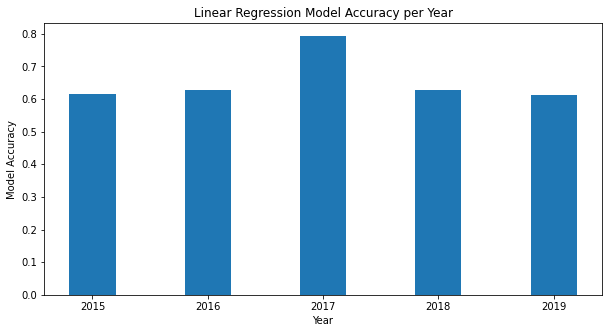
To construct our linear regression model, we mainly made use of the sklearn library from which we used the standard scalar, train test split and the linear regression extensions. For our train test split, we trained 80% of the data for each year and tested the model on the remaining 20%. Finally, we were able to find the predicted values for our test data and compare them to the actual values.

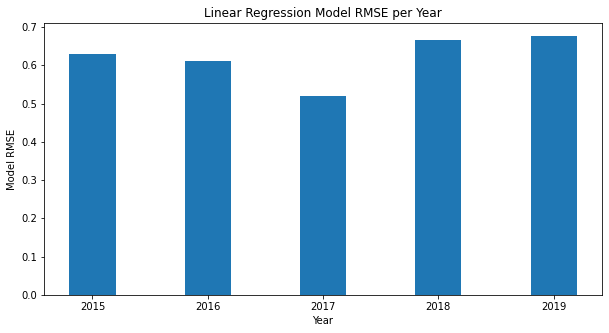
**Results**

KNN Classification Model

The KNN model gave us various results of accuracy after analyzing the dataset.. Using KNN, we were able to evaluate the successfulness of the model for each region identified shown below. The results show that the model was most successful with the region Western Europe. On the other hand, the model accuracy was relatively low for the regions Eastern Asia and Central & Eastern Europe. The results also show that the training accuracy for the model is about 78%. 

Linear Regression Model

After producing our linear regression models for each year, we were able to gather the predicted values of happiness score for 2015-2019. At first glance at the first few samples from each year, the predicted and actual values seem to be fairly close. However, in order to test our accuracy statistically, we scored the model for each year. The accuracies for the years 2015-2019 were approximately 0.614, 0.629, 0.793, 0.628, 0.613, respectively. A bar graph of these accuracies is provided below. 

Additionally, to understand how much error existed in our models, we calculated the root mean squared error for each year. These values were approximately 0.628, 0.611, 0.520, 0.666, 0.677 for 2015-2019, respectively. A bar graph for the error is also shown below.

**Conclusions**

As our team constructed two different types of models, our conclusions from this project are comprehensive and multifaceted. For our KNN classification model, we saw an overall accuracy of around 80%. Furthermore, the KNN model gave us consistent results and data for the precision and accuracy of each region. We can conclude that the fitting for the 2016 data after fixing the imbalance for the data can be fairly accurate at predicting the region using the countries happiness score and other characteristics. One thing we can change next time is to use machine learning in order to find the optimal K value instead of doing the rule of thumb of square rooting the number of samples to find it. This way it would be a more efficient method of finding the best K value to use for our KNN classifier model.

From the linear regression models, we mostly saw very consistent results over the five years. However, it is clear from our metrics of success that our 2017 model was the best out of all five years as it had the highest accuracy and the lowest root mean square error. In the future, we would be interested in further exploring the data to determine why the model of this year is an apparent outlier in accuracy. Nevertheless, from the results of all five years, our linear regression models were generally fair at predicting happiness scores.

The results from both models seem to suggest that the parameters of economy, life expectancy, freedom to make choices, and generosity were decent indicators of the happiness of people in a country. However, from this project we also learned that the happiness of individuals cannot be measured or predicted by just a few metrics. In the future, we hope that we have even more extensive data with more parameters that could affect the happiness of citizens of different countries.

**Works Cited**

Ologunsua, Success. "Is Happiness Free?" *Medium*, 23 Aug. 2020, medium.com/@TheSussex/is-happiness-free-bb9ef70d0bd5. Accessed 12 Dec. 2022.

Satadru5. "World Happiness Report Analysis." *Kaggle: Your Machine Learning and Data Science Community*, 21 Apr. 2017, www.kaggle.com/code/satadru5/world-happiness-report-analysis. Accessed 12 Dec. 2022.

Shinde, Manali. "Simple Applications of Multiple Regression Models — Part 2: Can We Predict World Happiness?" *Medium*, 11 June 2018, medium.com/one-datum-at-a-time/simple-applications-of-multiple-regression-models-part-2-can-we-predict-world-happiness-5c8a083e4e5e. Accessed 12 Dec. 2022.

Thorson, Kurt. "K-Nearest Neighbors — World Happiness Index 2021." *Medium*, Medium, 8 July 2021, k-thorson.medium.com/k-nearest-neighbors-world-happiness-index-2021-96a6f4b2e70. Accessed 12 Dec. 2022.

"Is There a Library Function for Root Mean Square Error (RMSE) in Python?" *Stack Overflow*, stackoverflow.com/questions/17197492/is-there-a-library-function-for-root-mean-square-error-rmse-in-python. Accessed 12 Dec. 2022.

Jeeteshgavande30. "Bar Plot in Matplotlib." *GeeksforGeeks*, 4 Mar. 2021, www.geeksforgeeks.org/bar-plot-in-matplotlib/. Accessed 12 Dec. 2022.

"Seaborn.heatmap — Seaborn 0.12.1 Documentation." *Seaborn: Statistical Data Visualization — Seaborn 0.12.1 Documentation*, seaborn.pydata.org/generated/seaborn.heatmap.html. Accessed 12 Dec. 2022.

Band, Amey. “How to Find the Optimal Value of K in KNN? - Towards Data Science.” *Medium*, 18 Apr. 2022, towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb.