WORLD HAPPINESS REPORT

DATA MINING FINAL PROJECT

Kıymet Deren Toy - Görkem Savran
kiymetderentoy@posta.mu.edu.tr - gorkemsavran@posta.mu.edu.tr

January 21, 2021

1 Introduction

First of all, our data set includes the names of countries, their happiness scores and the factors that affect this score. The source of our data is Kaggle and we have data for 5 different years as 2015,2016,2017,2018 and 2019. The general purpose of our project is to find out which factors affect the happiness score the most, to better understand the data by visualizing the data set, to cluster the countries according to the welfare level, to make a prediction about the next years.

2 Data Preprocessing & Visualizing

First, we loaded our 2019 dataset and checked the empty data. We did not do any preprocessing because there is no null data. Then we calculated the binary correlation of the columns and visualized them using a heatmap. Thus, we observed the relationships between the features. You can see output in Figure 2. Also, we had the pairplot plotted the pairwise relationships in the data set (Figure 3). Using the Plotly library, we showed the rankings of the countries on the map using the region information in our dataset (Figure 4). After these visualizations, we selected 7 countries from our data set based on their happiness scores. The top 3 of these countries were chosen from the happiest, 2 from the middle and the last 2 from the unhappy countries. We then compared these countries and used visualization to fully see the factors that made them happy or unhappy (Figure 5). Lastly, we visualized the changes in happiness scores for 5 years (Figure 6).

2.1 Backward Elimination

We used the backward elimination method to find the most important features that affect the score. We checked the P values of the columns by printing the OLS Regression Result. We proceeded by eliminating columns with a P value greater than 0.05. You can see the last output of this OLS report in Figure 7. As a result of this report, we found that the most influencing features were GDP per capita, social support and healhty life expectancy. And we verified this by visualizing the features that affect the score the most and those that affect the least (Figure 8).

3 Linear Regression

3.1 Predict 2019 Data From 2018 Data

In this section, we used the 2018 and 2019 data sets together. We filled the empty data using SimpleImputer because there was missing data in the 2018 dataset. Later, we trained a linear model with 2018 data. And we predicted the 2019 data and calculated the value of r2 score.

```
R2 score:  0.7755642097442412
```

Figure 1: R2 Score

3.2 Showing how GDP impacts by years

In addition to the linear regression above, we found how the common GDP per capita feature changes over the years in our 2015, 2016, 2017, 2018, 2019 data. We made a visualization by taking the weight of the GDP column each year (Figure 9).

4 Clustering

4.1 Visualizing Clusters

First, we did clustering with the Kmeans algorithm. We divided our dataset into 3 classes and visualized it (Figure 10).

4.2 Clustering Map

We clustered our 2019 dataset using sklearn.cluster and plotly.graph_objs. We created 3 clusters, they represented countries with happy, moderately happy and less happy. And we visualized them on the map (Figure 11).

5 Ensemble Learning

Previously, we had found the 3 most impressive features and visualized them. According to these 3 features, we subjected the 2018 and 2019 data to separately clustering and divided them into 3 classes. These 3 classes became our new labels. By combining 2018 and 2019 data, we applied various ensemble learning methods. These methods are briefly as follows: Boosting, voting and bagging. In Boosting, we trained our model with 10 different AdaBoost objects with estimators from 1 to 10. We created different MLP objects with hidden layers in Voting and added them to our VotingClassifier. We then trained our voting estimators and found the score for our VotingClassifier. Finally, we created a BaggingClassifier object with an estimator number of 10 and we trained our model with it. You can see the outputs in Figure 12.

6 Results

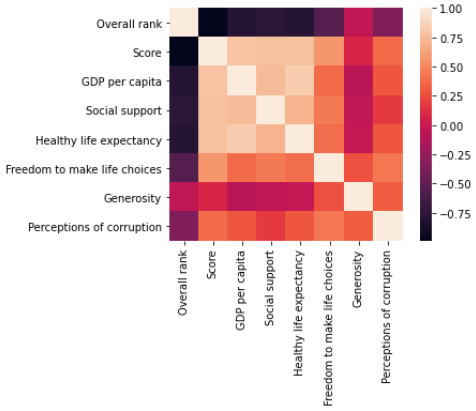

Figure 2: Heatmap

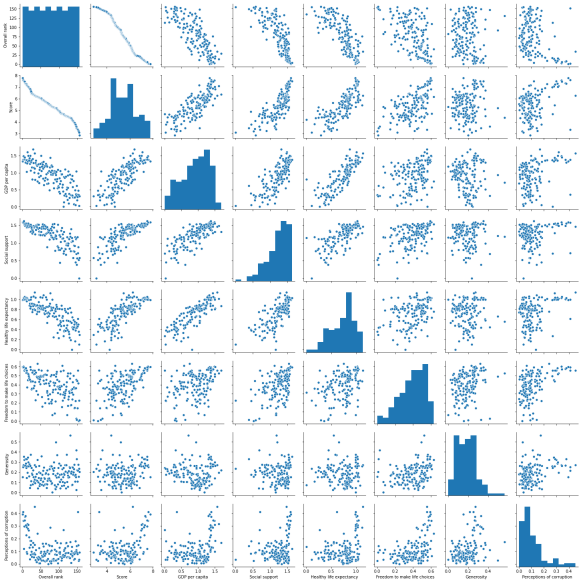

Figure 3: Plotting pairwise relationships in the dataset

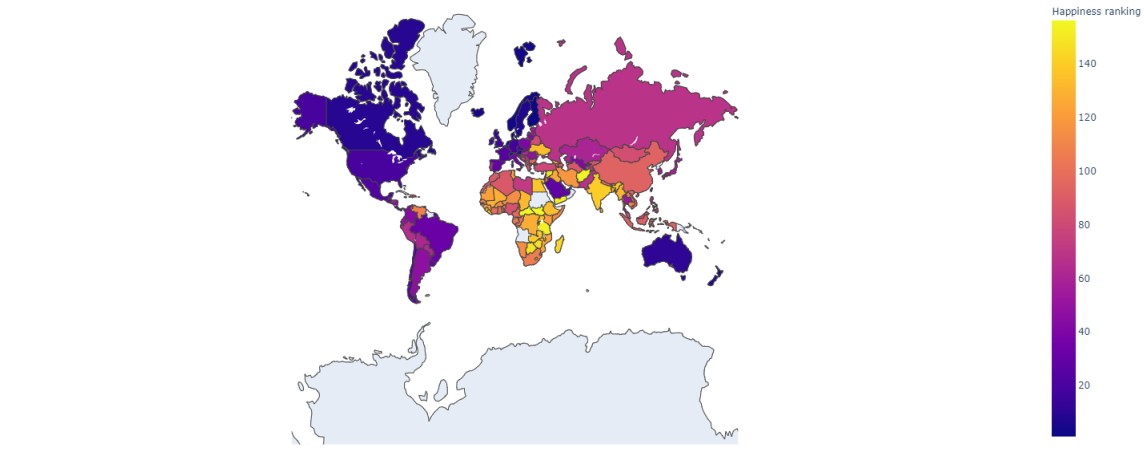

Figure 4: Happiness ranking of countries

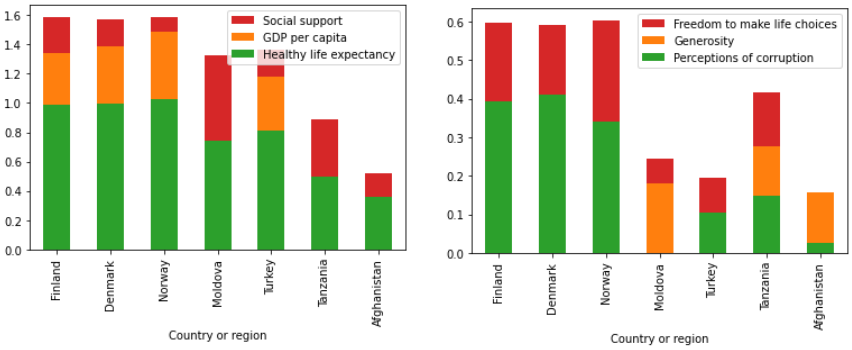

Figure 5: Comparison of countries according to 6 features

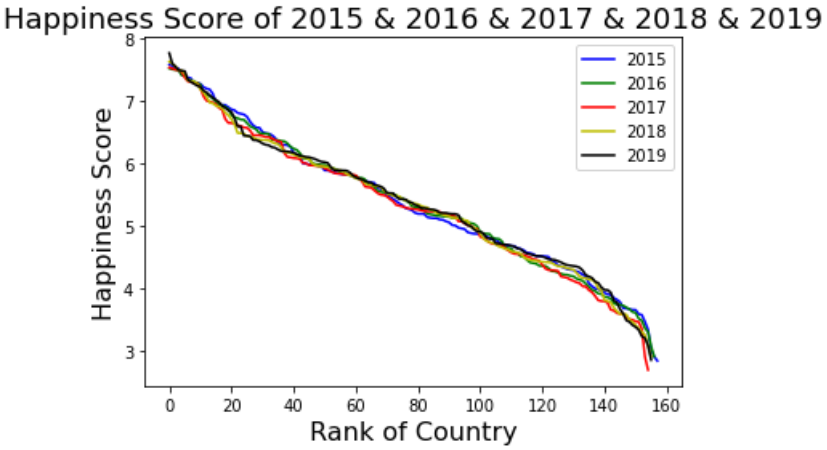

Figure 6: Happiness Score 2015&2016&2017&2018%2019

OLS Regression Results

Dep. Variable:	Score	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.765
Method:	Least Squares	F-statistic:	127.0
Date:	Wed, 20 Jan 2021	Prob (F-statistic):	2.82e-47
Time:	12:34:20	Log-Likelihood:	-122.62
No. Observations:	156	AIC:	255.2
Df Residuals:	151	BIC:	270.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8921	0.199	9.491	0.000	1.498	2.286
x1	0.8105	0.216	3.745	0.000	0.383	1.238
x2	1.0166	0.235	4.331	0.000	0.553	1.480
x3	1.1414	0.337	3.384	0.001	0.475	1.808
x4	1.8458	0.340	5.423	0.000	1.173	2.518

Omnibus:	5.077	Durbin-Watson:	1.641
Prob(Omnibus):	0.079	Jarque-Bera (JB):	4.685
Skew:	-0.413	Prob(JB):	0.0961
Kurtosis:	3.198	Cond. No.	17.8

Figure 7: Last output of OLS Regression Results

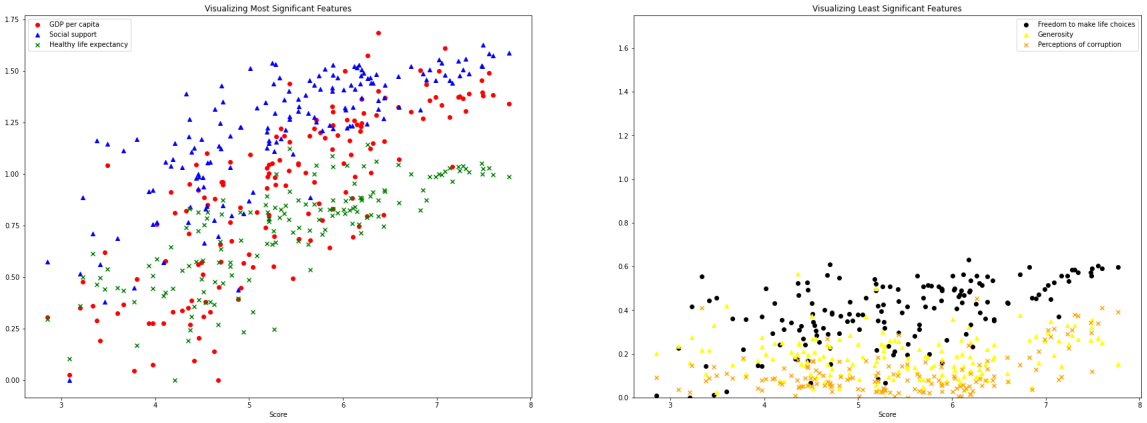

Figure 8: Visualizing Features

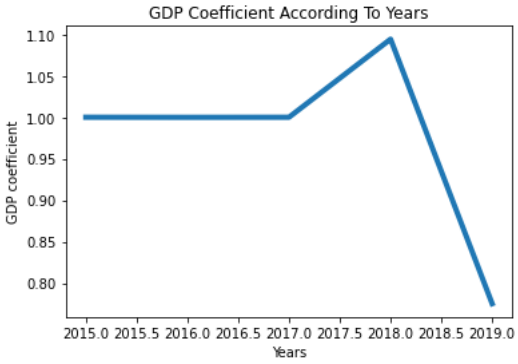

Figure 9: GDP coefficient according to years

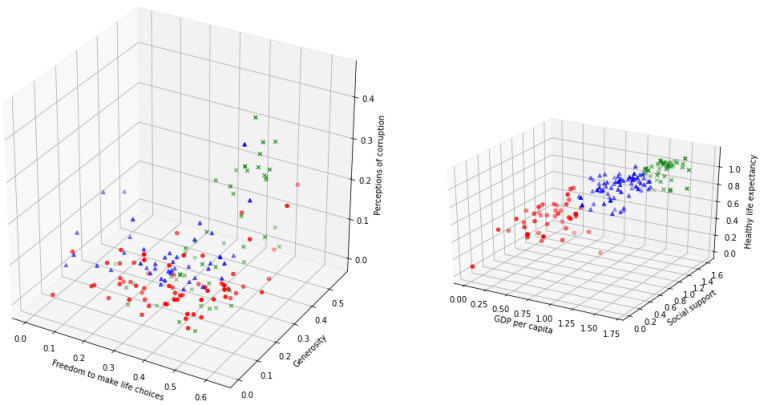

Figure 10: Clustering

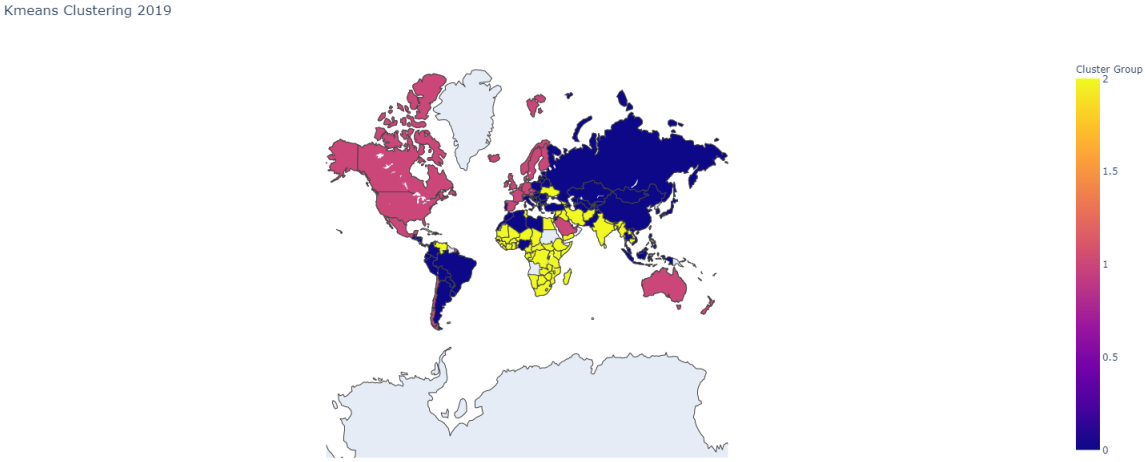

Figure 11: Kmeans Clustering 2019

```
(1 Estimators)AdaBoost Score: 0.908256880733945
(2 Estimators)AdaBoost Score: 0.7981651376146789
(3 Estimators)AdaBoost Score: 0.944954128440367
(4 Estimators)AdaBoost Score: 0.7981651376146789
(5 Estimators)AdaBoost Score: 0.9678899082568807
(6 Estimators)AdaBoost Score: 0.7981651376146789
(7 Estimators)AdaBoost Score: 0.9541284403669725
(8 Estimators)AdaBoost Score: 0.9678899082568807
(9 Estimators)AdaBoost Score: 0.9724770642201835
--------------------------------------------------
Voting Classifier with different MLPs Accuracy: 0.6808510638297872
Bagging Score: 0.9574468085106383
```

Figure 12: Ensemble learning scores

7 Conclusion

As a result, using Figure 2, Figure 3, and Figure 8, we showed that the features that most affected our data set were GDP per capita, social support, and healthy life expectancy. We checked and verified this information statistically using the backward elimination method. In other words, we have revealed the country characteristics that are necessary for people in a country to be happy and to live a good life.

Apart from that, in Figure 4, we visualized our data set on the map so that the happiness rankings of the countries appear as a whole and our data can be understood more easily.

Later, in Figure 5, we compared the 3 happiest countries with the moderate and less happy countries, and as a result of this comparison, we saw that the most important feature in the 3 happiest countries was healthy life expectancy. In addition, we observed that the least common feature in these 3 countries is generosity. The common feature of the most unhappy countries was that GDP per capita values were very low. The countries which have middle happines score (such as Turkey) might seem that their most important features are good, but the low scores in the least affecting features has led to remain in their mid status.

In Figure 6, we have shown that the happiness score corresponding to each country's rankings has not changed much over the years.

In the linear regression part of our project, we first estimated the 2019 data set from the 2018 data set. We used the R2 score to measure the accuracy of our model. As seen in Figure 1, this value came around "0.77". This showed us the accuracy of our model. (Note: The value of R2 when the model predicts 100% correct is 1.) This trained model can also be used to predict future years.

In the second part on linear regression, we visualized the effect of GDP per capita over all years. As can be seen from Figure 9, while GDP per capita value was given a stable importance before 2018, in 2018, a lot of attention was paid to GDP per capita value. In 2019, this importance has decreased even more than it was in 2015, 2016, 2017. This has shown us that the factors that people care about may change over the years and lead to changes in the ranking of happiness.

Regarding the clustering part, we first clustered our data set in Figure 10 according to the most effective and least effective features. As a result, we found that GDP per capita, healthy life expectancy, social support features were in a linear correlation. We have seen that other features are not correlated in the same way and are irregular in such a way that no meaning can be made.

Finally, we have shown the countries in 3 different groups by using visualization on the map again in Figure 11. The first of these groups belongs to the happiest countries, the other to the countries of medium happiness, and the last to the unhappy countries. Thanks to this visualization, we analyzed our data more easily, showed in which continent there are more happy people and the welfare level of the countries.

Some of the models we trained with the AdaBoost objects we created had scores of 95% and some 79%. Our model, which we trained with Voting Classifier, or MLPs, showed a worse result compared to other algorithms and remained around 68%. Our model, which we trained with bagging, showed a very good success with a score of 95%. Considering these results, we have seen that it would be better to use Boosting or Bagging when such data is encountered. Thanks to these algorithms, we achieved very good results using weak learners and we did unsupervised learning by creating our own classes with clustering.