Machine Learning Report

Diego Quintana

Marcel Pons

Manuel Breve

Machine Learning Course
Master in Innovation and Research in Informatics

Universitat Politècnica de Catalunya

Facultat d'Informatica de Barcelona

- June 2020 -

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

FIB

Machine Learning Project

World Happiness Dataset

The World Happiness Report[1] is a landmark survey of the state of global happiness that ranks, through a Happiness Score, 156 countries by how happy their citizens perceive themselves to be.

In this project, we analyze the World Happiness Report for the years 2018 and 2019 through different machine learning techniques. First, during the exploration stage of the data, we found that the GDP per capita per country is highly correlated with the Happiness Score reported, and that we can efficiently cluster them into three groups.

Finally we built different machine learning models, both linear and nonlinear, such as linear and polynomial regression, decision trees, random forests and neural networks in order to predict the 2019 Happiness Score, being neural networks the one with the best performance, i.e, the minimum RMSE value.

The World Happiness Dataset is made of one (1) continuous response variable, six (6) continuous independent variables, and one (1) supplementary categorical variable.

Continuous Response Variable:

1. **Score**: This Happiness Score is a subjective well-being perception based on the survey's answers, which ask people to evaluate different subjects about their life quality on a scale of 0 to 10.

Continuous Independent Variables:

Among the survey variables, we have selected the following six continuous variables in order to make the multivariate analysis:

1. **GDP per capita**: in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank. (Using natural log of GDP per capita, as this form fits the data significantly better than GDP per capita)
2. **Social support**: the national average of the binary response (either 0 or 1) to the Gallup World Poll (GWP) question "If you are in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?
3. **Healthy life expectancy at birth**: constructed from the World Health Organization (WHO) and WDI. Adjustment by applying the country-specific ratios to other years.
4. **Freedom to make life choices**: the national average of binary response to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
5. **Generosity**: residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.

[1] https://worldhappiness.report/

6. **Perceptions of corruption**: the average of binary answers to two GWP questions: “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?”. Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

Supplementary categorical variable:

1. **Regional Indicator:** Region in the world, made of 10 levels: Western Europe, Central and Eastern Europe, Sub-Saharan Africa, Middle East and North Africa, East Asia, Southeast Asia, South Asia, North America and ANZ, Latin America, Caribbean and Commonwealth of Independent States.

Overall rank	Country	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Regional Indicator
1	Finland	7.632	1.305	1.592	0.874	0.681	0.202	0.393	Western Europe
2	Norway	7.594	1.456	1.582	0.861	0.686	0.286	0.34	Western Europe
3	Denmark	7.555	1.351	1.59	0.868	0.683	0.284	0.408	Western Europe
4	Iceland	7.495	1.343	1.644	0.914	0.677	0.353	0.138	Western Europe
5	Switzerland	7.487	1.42	1.549	0.927	0.66	0.256	0.357	Western Europe

Finland, Norway, Denmark, Iceland, and Switzerland are the countries with the highest Happiness Score

Data Exploration

Pre-processing

From the summary of the dataset we can see that *Overall.rank* represents the Happiest ranking in terms of the Score, so we have proceeded to remove it because it produces redundant information.

We have created two versions of the train dataset. The former includes the Regional.Indicator to be used in terms of visualizations, while the latter does not contain this variable in order to work with numerical variables only.

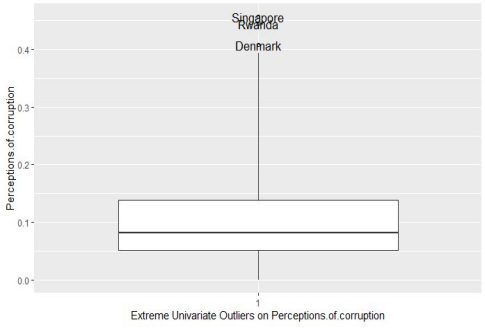
Missing Values

There exists one missing value on the Perceptions.of.corruption variable in the train dataset, the one has been imputed.

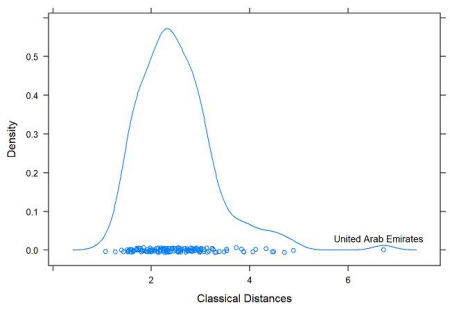
Outliers

We have also detected four outliers. Three of them: Denmark, Singapore, and Rwanda, were detected as univariate outliers in terms of the Perceptions.of.corruption variable. While

the fourth: United Arab Emirates was considered an outlier given the Classical Mahalanobis distance.



Perceptions.of.corruption Boxplot

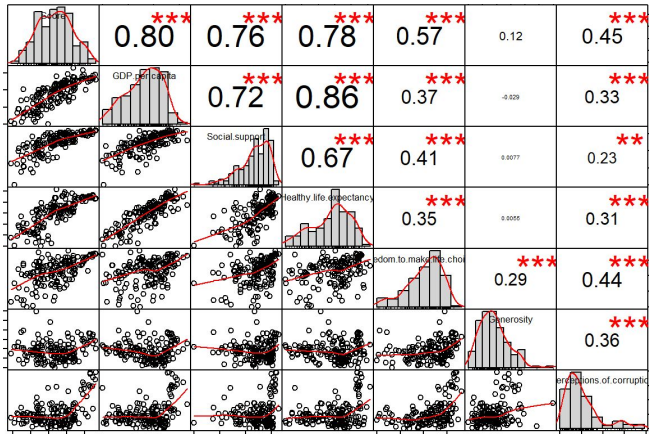


United Arab Emirates as Multivariate Outlier

Feature Selection

We would like to take a last look at our variables in order to select those that help us to reduce overfitting, improve accuracy, or reduce training time.

We observed that GDP.per.capita is the variable with highest correlation with Happiness Score with a value of 0.80 in Pearson's correlation, followed by *Health.life.expectancy*, so countries with a higher GDP per capita seem to be happier, yet this time we have decided to maintain our actual variables and finish our pre-processing.

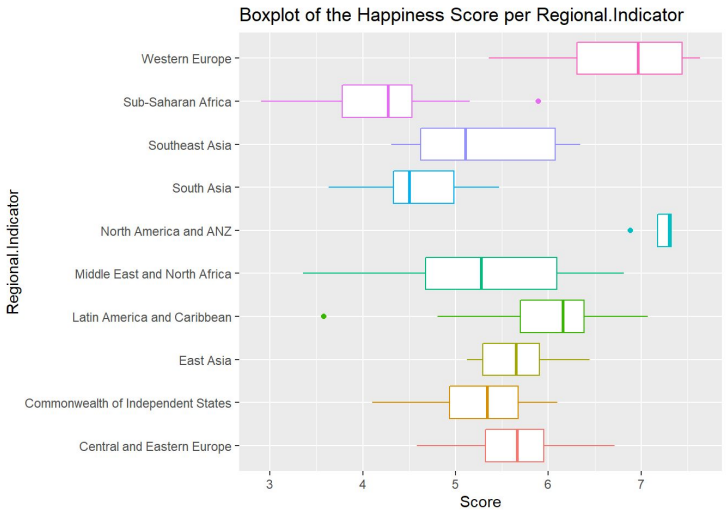


Correlations between variables

Visualization

To continue with our data exploration, this time we would like to create and visualize different clusters of countries according to the available variables.

First, we can take a look of the Happiness Score per Regional.Indicator, where we can observe that the North America and ANZ are the most happiest countries in the World, followed by Western Europe and Latin America and Caribbean, while the least happiest countries are located in Sub-Saharan Africa Region.

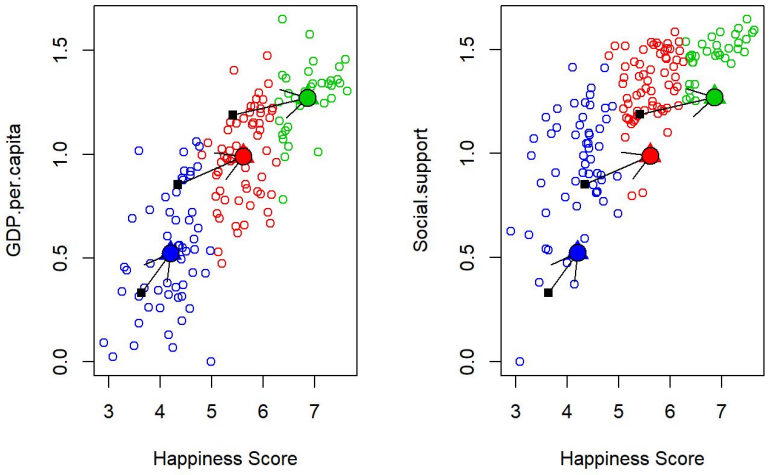


North America and ANZ it's the Region with the highest Happiness Score

K-means

Now, we would like to create new clusters, different from the Regional.Indicator, so this time we're working with our numerical matrix in order to discover new groups of countries.

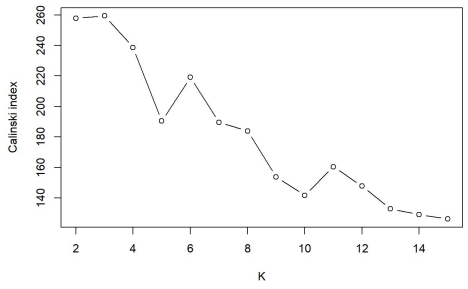
First, we try a K-means clustering technique with K=3. We can see the change in the coordinates of the centroids, from their first position with black squares to their final position in colored circles. We can clearly see the correlations of some variables when we see the clusters of countries grouped to each other.



K-means clustering, K =3

Clustering quality can be measured by the Calinski-Harabasz index, so we would like to know if K=3 is the best parameter to our dataset.

From the plot, we observe that the highest, i.e, the best Calinski-Harabasz index value corresponds to the K=3.



K =3 has the biggest Calinski-Harabasz index for clustering

PCA Interpretation with Dimensionality Reduction

Sometimes it can be messy to deal with a large number of variables, so we have reduced the dimensionality of our dataset through PCA.

	Dim.1 <dbl>	Dim.2 <dbl>
Finland	3.401397	1.4063637
Norway	3.469523	1.6253963
Iceland	2.805805	0.9264636

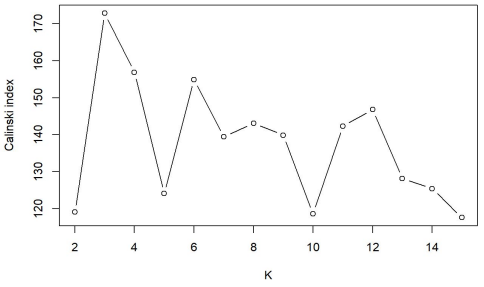
New dataset configuration on two dimensions after PCA

The seven original variables can be reduced to 2 of them, holding for ~73% of the original information.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.9756223	49.593706	49.59371
comp 2	1.4018199	23.363666	72.95737
comp 3	0.5878915	9.798192	82.75556
comp 4	0.5756055	9.593425	92.34899
comp 5	0.3281903	5.469839	97.81883
comp 6	0.1308704	2.181173	100.00000

First two Eigenvalues retain ~73% of the original information

We have repeated the K-means grid-search in order to check if now there exists a new optimum K, yet K=3 is once again the optimal value for clustering.

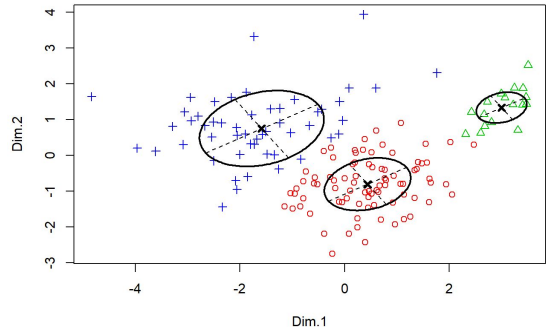


K =3 has the biggest Calinski-Harabasz index for clustering with new dataset dimensions

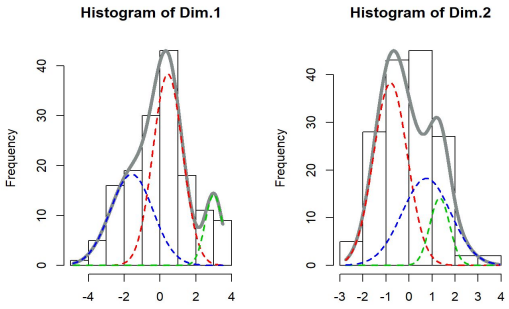
E-M Clustering

We can also use this modified training dataset with two dimensions to run another kind of clustering technique. EM Clustering is similar to the K-Means technique, but its main goal is to maximize the overall probability or likelihood of the data given the clusters, so it allows us to compute new and interesting plots, as the possible former gaussians of the countries.

We ran a 10 clusters grid search, getting K=3 as the optimal cluster numbers.



3 Clusters and their centroids using E-M Clustering



Clusters Distribution on two dimensions

Machine Learning Modeling

Machine learning methods allow us to create different predictive models, yet choosing the appropriate one it's not always a trivial task. Looking for the best model we have considered different options, including linear and non-linear methods, such as linear and polynomial regressions, decision trees, random forest, and neural networks.

The different models were built with the 2018 Happiness Ranking dataset and were first evaluated using different measurement techniques. In regression we run a 10-cross fold validation resampling procedure over the training dataset, the one was splitted 70%-30 for train and test, evaluating and comparing the regression models through the Normalized Mean Squared Error, NMSE. In decision trees we have also used a cross-validation error function, while in random forests we have the OOB estimate of error rate. Finally, in neural networks we ran a 10-cross fold validation and evaluated the Root Mean Squared Error, RMSE.

Having different measurement procedures for the models makes it difficult to fairly compare their performance, so the best model has been chosen through the prediction error for the 2019 Happiness Score values and comparing them with their real values by means of the Root Mean Squared Error, RMSE.

We observed that different runs of neural networks for training may generate different solutions. In short, in some runs the neural networks were the best models, narrowly overcoming random forests, yet sometimes their performance wasn't that good, being random forests the best model. In terms of the regressions, the linear standard regression was the one with better performance. Decision trees on the other side did not perform as expected.

1. Regression models

To evaluate the different regression models we selected the best parameterization for linear regression, ridge and lasso regularized linear regressions, and polynomial regression, and then ran a 10 cross fold validation in order to compute NMSE. We describe how we got the best parameters for different regression models and then we compare their results.

1.1. Standard Linear Regression

A model with the `lm()` function was built, on which a backward stepwise feature selection was performed in order to find not useful variables for the model.

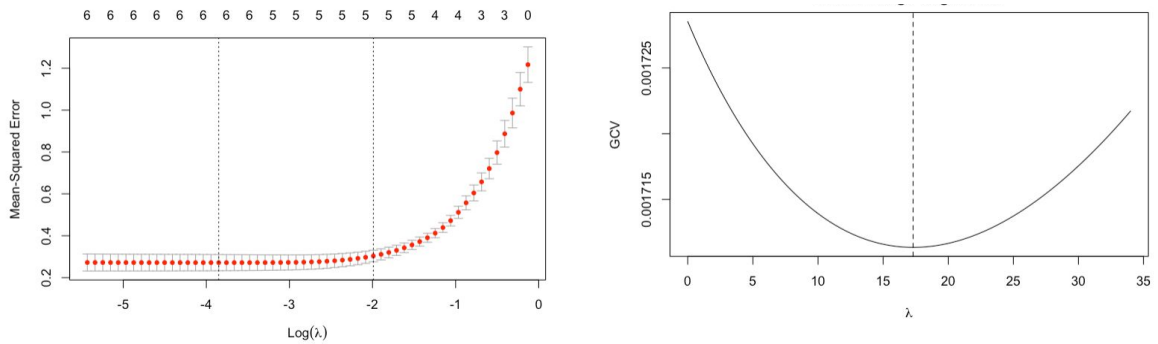
From our data, the only variable that happened to reduce the AIC was Generosity, which was also the only variable that was not significantly correlated with the Happiness Score (with a Pearson's correlation of 0.12). Therefore, when comparing models, a standard linear regression without this variable was performed.

1.2. Ridge Linear Regression

Ridge regression is a regularized regression with the coefficients penalized by the L2 norm. In order to find the optimal regularization parameter λ, the function `lm.ridge()` has the built in cross validation option to find the best one. The best λ for this model is 17.3.

1.3. Lasso Linear Regression

In Lasso regression, the coefficients are penalized by the L1 norm and for this reason, unlike ridge regression, the slope of the coefficients can reach 0 and consequently some variables can be excluded from the model. When the Lasso regression was built with the `glmnet()` function, considering all variables, the model did not take into account the Generosity variable, coinciding with the results of the stepwise feature selection. Like ridge, the optimal value for λ was chosen by cross-validation, which in this model was 0.02126.



Minimum λ for the regularized linear models Lasso (left) and ridge (right) regression found by 10-Fold Cross Validation

1.4. Polynomial Regression

In polynomial regression, the best n^{th} degree polynomial in x that best fitted our data was found by means of 10-fold cross validation. The degrees from 2 to 10 were considered, being the $6t^{h}$ degree the one that gave the minimum cross validation error.

1.5. Regression comparison

To select the best model, a 10-fold cross validation was performed, using as λ's in ridge and lasso the ones found previously, the 6^{th} degree in the polynomial regression, and removing the Generosity variable in the standard linear regression.

	NMSE train <dbl>
Standard	0.2208107
Ridge	0.2214507
Lasso	0.2243004
Poly	0.3184740

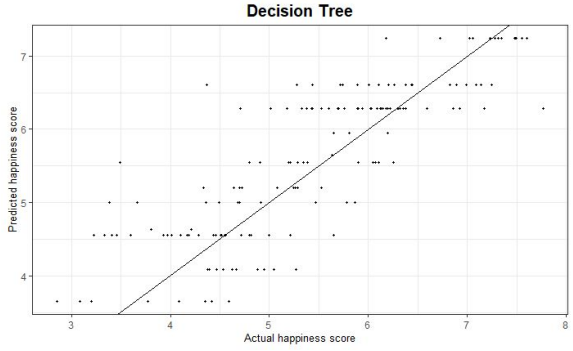
NMSE values in different regression types

We chose the best model as the one with the lowest CV error, which turns out to be the standard linear regression.

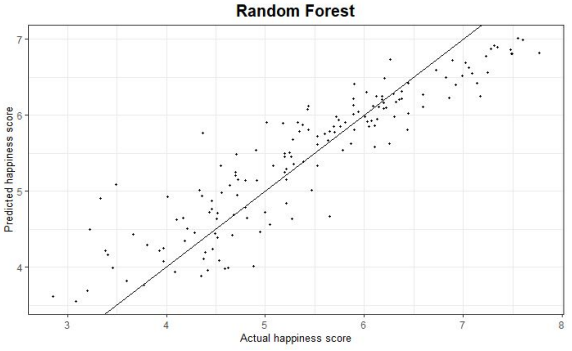
2. Decision Trees

Decision trees are easy to interpret and understand, yet their performance could be far from expected. Random forests on the other hand combine multiple decision trees, so it becomes more difficult to interpret, yet they obtained a great performance when predicting the 2019 World Happiness Score.

When testing the models with the 2019 World Ranking Happiness dataset we can observe how far are the data points respect to their real values on the decision trees models, while Random forests kind of approximate to a straight line.



Error on decision tree



Error on random forest

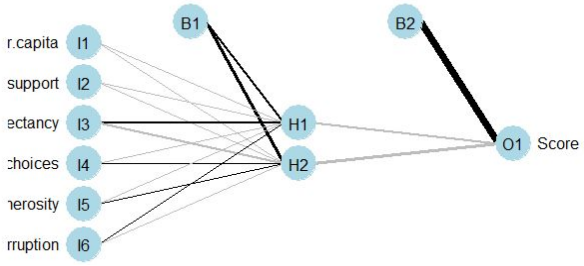
3. Neural Networks

Neural networks provide a non-linear method to predict new scores. One important thing about neural networks it's about their hyper-parametrization regarding the number of hidden layers, and its size.

For the purpose of our investigation, we tried with two configurations over a single hidden layer neural network. In the first configuration (nn1), we run a grid search looking for the optimal size of the hidden layer, which it was two. The second configuration (nn2) was another grid search trying different decay values in order to avoid over-fitting.

For both configurations we built the models pre-scaling the data and evaluating them through 10 cross fold validation with the 2018 Happiness dataset.

Different runs of the neural networks turned into different results, yet our first configuration (nn1) with one hidden layer of size two and no regularization turned out to be the best model, i.e, the one with the smaller error (RMSE).



Neural Network with one hidden layer of size 2

Final Model

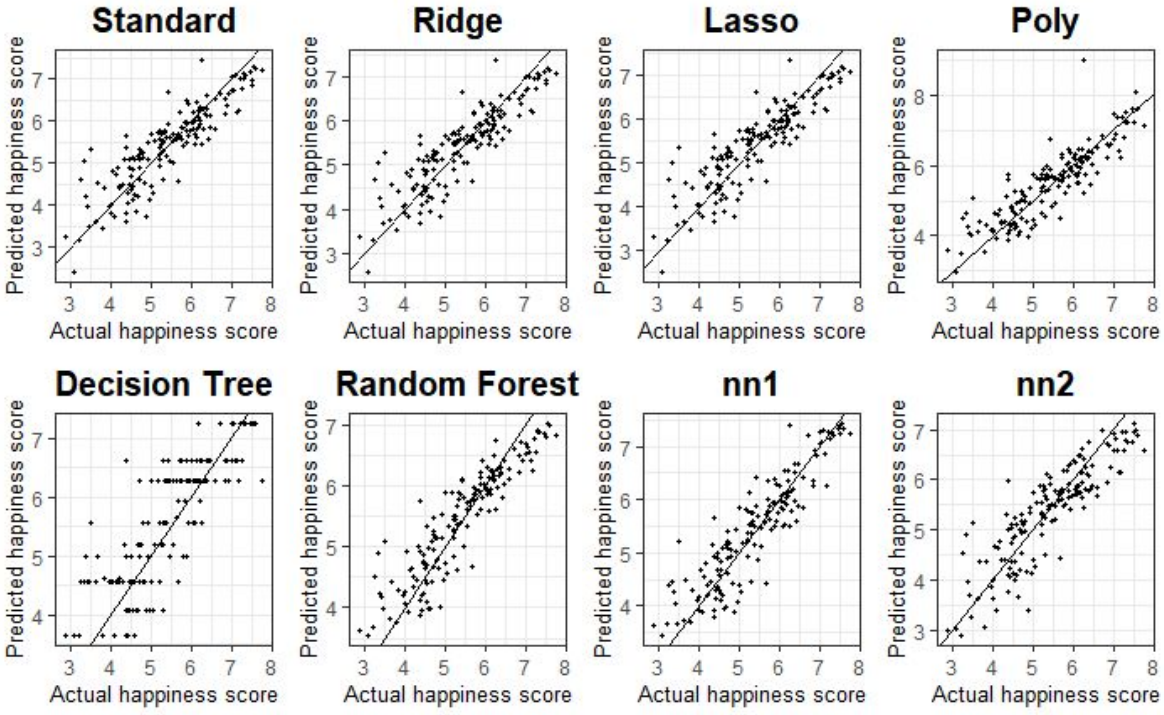
In order to select the best model we have evaluated their performance using the 2019 World Ranking Happiness Score and compare the predictions with the real values through Root Mean Squared Error, RMSE.

	RMSE <dbl>
Standard	0.5253792
Ridge	0.5255255
Lasso	0.5249096
Polynomial	0.5406458
decisionTree	0.6483592
randomForest	0.4829486
nn1	0.4818080
nn2	0.5437466

RMSE value for different machine learning models with 2019 World Happiness dataset

So far, neural networks and random forest were the best models, yet neural networks performed a little better, with 0.4818 RMSE value.

It is also interesting to visualize the plots of the actual vs predicted happiness scores of the 2019 dataset, where we can observe how the NN1 model has the better fit, while in the decision tree model the predicted scores are far away from their real values.



Actual vs Predicted values on different machine learning models

Conclusions.

We have studied the World Happiness Report through different Machine Learning techniques. It is interesting to analyze the countries of the world through such relevant variables as GDP.per.capita, which it turned to be one of the most correlated variables with the national Happiness Score.

We can also visualize and cluster the countries in different groups. Three clusters of countries were clearly defined and segmented by some correlated variables as their GDP.per.capita.

Regarding the machine learning methods we have observed the power and flexibility of non-linear models such as neural networks random forests. While neural networks are often excellent choices as a model, it is still important to know how they work and be minded that different runs can produce different results, and some of them didn't were better than random forest, which is less computationally expensive than neural networks. So far, our dataset was a little small and neural networks will require more data to build a robust model.

It will be interesting to try and build new models in the future in order to predict the World Happiness Score in order to help in building a better and happier society.