CS4038 Data Mining Assignment

Happiness Score Predictor

Submitted by

Akhil Joseph Sunny B160470CS
Akshaya Kumar B160852CS
Mathew George B160008S
Pavan Kalyan B160723CS

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

March 14, 2019

Happiness Score Predictor

The objective of the assignment was to analyze various world development indicators to predict the happiness score of a country. To accomplish this, the World Happiness Dataset for the years 2015-2017 from Kaggle and the World Development Indicators from World Bank website were used.

The World Happiness Dataset consisted of Happiness score which was divided into Economy, Family, Health, Freedom, Trust, Generosity and Dystopia. The World Happiness Dataset was reduced to represent Country name/code,Year and Happiness Score.33 World Development Indicators from the respective dataset were used.

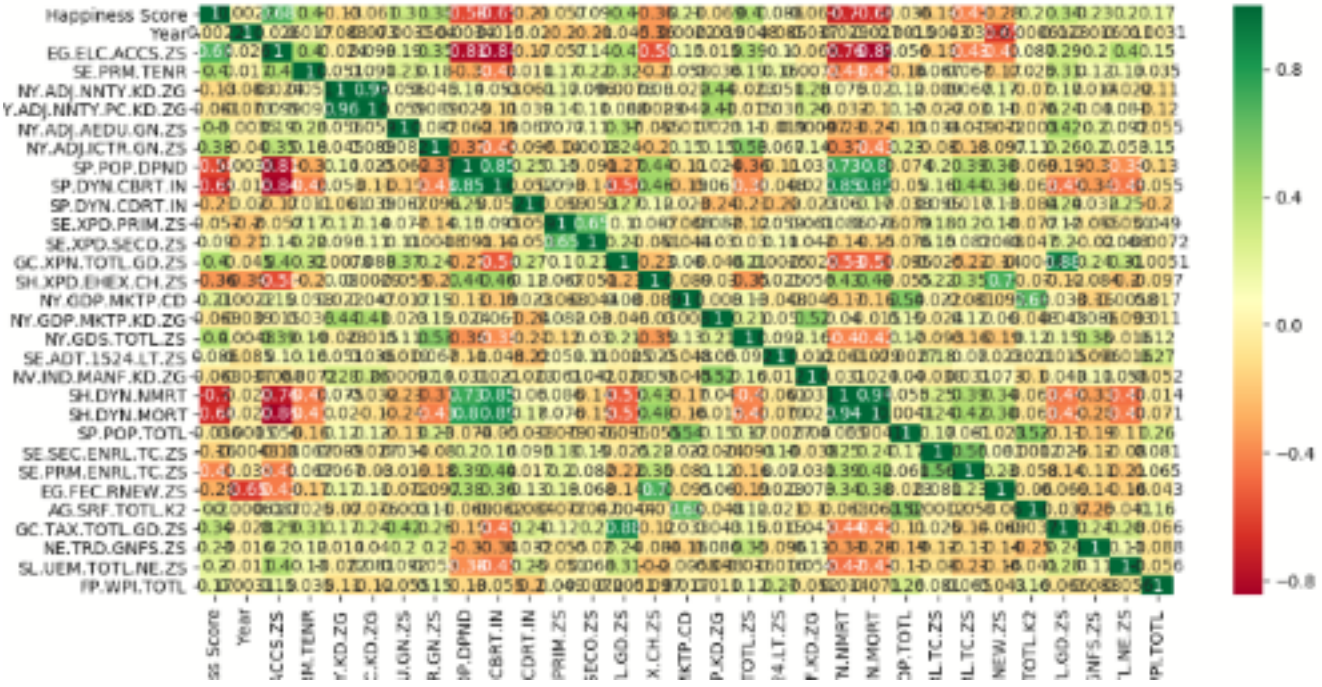
The Data Processing stage involved merging of all the above the datasets based on country name and year. The final dataset contained 414 values. The missing values were filled using mean of the values of other years.

A correlation heat-map was plotted to understand the relationship between the various attributes in the Data Analysis stage. Also, graphs were plotted to visualize the correlation between each of the indicators. The following indicators were removed due to presence of excessive null values.
1.per_sa_allsa.cov_pop_tot
Coverage of social safety net programs (% of population)
2.per_allsp.cov_pop_tot
Coverage of social protection and labor programs (% of population)
3.per_si_allsi.cov_q5_tot
Coverage of social insurance programs in richest quintile (% of population)
4.per_si_allsi.cov_q1_tot
Coverage of social insurance programs in poorest quintile (% of population)

The dataset was scaled and using the correlation coefficient in the previous step, the following features were extracted as part of Feature Selection:
1.EG.ELC.ACCS.ZS
2.SP.POP.TOTL
3.SE.XPD.PRIM.ZS
4.NV.IND.MANF.KD.ZG
5.NY.ADJ.NNTY.PC.KD.ZG

Different machine learning models were used for predicting the Happiness scores using the attributes selected from the previous stage. From the dataset, the rows corresponding to the years 2015-2016 were used as the training set and the year 2017 was used as the test set. The Linear Regression model had an accuracy of 77.14%. The Random Forest Regressor had an accuracy of 89.4%.

Conclusion
Happiness scores of various countries could be predicted with a fairly good accuracy using different data mining techniques. The model which gave the best result is the Random Forest Regression model.


Happiness Score
Year
EG.ELC.ACCS.ZS
SE.PRM.TENR
NY.ADJ.NNTY.KD.ZG
Y.ADJ.NNTY.PC.KD.ZG
NY.ADJ.AEDU.GN.ZS
NY.ADJ.ICTR.GN.ZS
SP.POP.DPND
SP.DYN.CBRT.IN
SP.DYN.CDRT.IN
SE.XPD.PRIM.ZS
SE.XPD.SECO.ZS
GC.XPN.TOTL.GD.ZS
SH.XPD.EHEX.CH.ZS
NY.GDP.MKTP.CD
NY.GDP.MKTP.KD.ZG
NY.GDS.TOTL.ZS
SE.ADT.1524.LT.ZS
NV.IND.MANF.KD.ZG
SH.DYN.NMRT
SH.DYN.MORT
SP.POP.TOTL
SE.SEC.ENRL.TC.ZS
SE.PRM.ENRL.TC.ZS
EG.FEC.RNEW.ZS
AG.SRF.TOTL.K2
GC.TAX.TOTL.GD.ZS
NE.TRD.GNFS.ZS
SL.UEM.TOTL.NE.ZS
FP.WPI.TOTL
0.8
0.4
0.0
−0.4
−0.8