CS4038 Data Mining Assignment

Happiness Predictor

Submitted by

Arun Joseph	**B150102CS**
Athul Ajesh	**B150167CS**
Vysakh D	**B150083CS**
Roshith Raghavan	**B110219CS**

तमसो मा ज्योतिर्गमय

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

March 1, 2019

Happiness Predcitor

Arun Joseph Athul Ajesh Vysakh D Roshith Raghavan

The assignment aimed at analyzing various world development indicators to predict the happiness score of a country. The World Happiness Dataset from Kaggle and the World Development Indicators from World Bank website were used as the datasets.

The World Happiness Dataset for each year contains the country name and happiness score. The Happiness score is divided into Economy, Family, Health, Freedom, Trust, Generosity and Dystopia. The World Happiness Dataset for 2015-2017 were used for this assignment. 33 indicators were picked from the World Development Indicators. In the Data Processing stage, all the above the datasets were merged on the basis of country name and year. A total of 131 countries were present in the final dataset. The missing values were filled using interpolation.

In the Data Analysis and Visualization stage, a correlation heat-map and matrix were plotted to visualize the relationship between the various attributes. Also, graphs were plotted to identify the correlation between each of the indicators. From the above methods, attributes relating to National Income, coverage of social programs, mortality rate and wholesale price were related or contained a lot of null values. Thus, these attributes were removed form the dataset. The attributes which had the highest correlation with the happiness score were access to electricity, birth rate, mortality rate and pupil-teacher ratio.

In the Feature Extraction stage, the dataset was normalised. Using Recursive Feature Extraction, Principal Component Analysis and the results from the previous stage, the attributes which gave better results were selected for each of the split-up scores by looking at the predictions using a Linear Regression model.

For the last stage, different machine learning models were used for predicting the split-up 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 happiness score was calculated by adding the split-up scores. In the Linear Regression model, the R2 score was 0.571 with an accuracy of 88%. The Random Forest Regressor gave the best performance with a R2 score of 0.757 and an accuracy of 91.45%.

In the future, for getting a better result, the attributes which were not used could be replaced with different attributes which could predict split-up scores like Family and Dystopia with a better accuracy. Another approach would be to add some of the split-up scores to compensate for the lack of attributes with a good correlation score.

As a conclusion, we were able to predict the happiness scores with a fairly good accuracy using different data mining techniques. The model which gave the best result is the Random Forest Regressor.

The code is available at: https://github.com/arun-Joseph/Happiness-Predictor.