World Happiness Ranking Prediction

Asiyah Zukfily, Athirah Zaini, Fatihah Zamri and Shazana Affandi

***Abstract* — World Happiness Report is a survey of the state of global happiness. The reports review the state of happiness around the world in 2016. This dataset gives the happiness rank and happiness score of 156 countries around the world based on six factors which is economic production, social support, life expectancy, freedom, absence of corruption, and generosity. Our objectives are to predict countries or regions rank the highest in overall happiness, to estimate the six factors contribute highest happiness and to observe any country experience a significant increase or decrease in happiness. This project produces a good result in finding out the major factors of overall happiness of that country. Besides, it able to acquire which countries ranked highest in overall happiness through this project.**

I. INTRODUCTION

The World Happiness Report is a survey of global happiness including all 156 countries in the world. The main purpose for this project is to know the scores and the ranking of happiness for each country in the world and to know the reasons behind it. The happiness scores and rankings use data from the Gallup World Poll where the scores are evaluated based on the answers of the citizens for each country to the main life evaluation question asked in the poll. Respondents are asked to answer the questions and rate their life from zero to ten which the best possible life for them is being a 10 and the worst possible life being a 0. In this project, there are six factors that are asked as the causes to the happiness of the citizen in each country which are economy, family, health, freedom, trust and generosity. These six factors will contribute to the scores and rankings whether the happiness for the specific country is higher or lower than others.

Given the World Happiness Report 2016, the dataset used contains all the variables needed to help us solve our problem and answer our questions regarding the scores and the ranking of happiness for all the countries in the world. Columns in the datasets, which are the six factors, plays the important role that can make life evaluations go higher in each country. To understand our dataset, we need to understand our variables.

Our datasets contain 13 variables and 157 instances. Our first variable is country which is nominal variable. It contains 156 countries. Our second variable is Region which is also nominal. This breaks the data up into regions according to continent. Our third variable is happiness rank which is numeric and ranks each country's happiness from 1-156 based on the world happiness score. The world happiness score variable is numeric and is based on a happiness scale from one to ten. Our lower and upper confidence levels are used to predict means these are both numeric. We also are given a GDP per capita numeric variable which gives us insight into how well a country is doing economically. Family is a numeric variable which gives us insight in how much a country values family. Health care is a numeric variable which is based on access, and quality of health care. Freedom and trust are a numeric variable. This is based on how free people feel in the given country and how much do they trust the government, Generosity is a numeric variable based on how generous the people are in the given country. These variables are used to compare the scores for each factor with other countries and evaluate their level of happiness. So, the ranking will be estimated through the level of their happiness.

II. PROPOSED MODEL

In our project, we used several operators that plays important role for achieve our goal which is to find what are the reason for the country to ranked highest in happiness. The operators we used is Select Attributes, Set Role, Optimize Selection, Cross Validation, Classification by Regression, SVM, Multiply, Random Forest, Apply Model, Decision Tree, Performance, Correlation Matrix and Aggregate. For the first part, we used select attributes to choose appropriate attributes in our dataset. We only select eight from thirteen attributes since another five attributes seems to be not important for our goal. The attributes selected is Country, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual.

Next, we set country as label in set role. Optimize selection in an operator to select most relevant attributes of the dataset. The optimize selection operator is applied on the data set which is a nested operator as it has a subprocess. It is necessary for the subprocess to deliver a performance vector. This performance vector is used by the underlying feature reduction algorithm. The cross validation operator has been used which itself is a nested operator. The subprocess of the cross validation contains training and testing part which is the classification by regression is used in training part while apply model which applying training model and performance is used in testing part. For the subprocess of classification by regression, we used SVM operator.

Correlation matrix is used to show whether and how strongly pairs of attributes are related. The Aggregate operator creates a new data set from the World Happiness 2016 data set showing the results of the selected aggregation functions. Many aggregation functions are supported including SUM, COUNT, MIN, MAX, AVERAGE and many other similar functions known from SQL. This operator performs the aggregation functions known from SQL. The functionality of the GROUP BY clause of SQL can be imitated by using the group by attributes parameter. Multiply is used to takes the RapidMiner Object from the input port and delivers copies of it to the output ports. Each connected port creates an independent copy. So, changing one copy has no effect on other copies.

The result we get from the World Happiness 2016 data set, optimize selection and correlation matrix is applied in random forest and decision tree operator. A random forest is an ensemble of a certain number of random trees, specified by the number of trees parameter. These trees are created on bootstrapped sub-sets of the data set provided from optimized selection. For classification, the rule is separating values belonging to different classes. A decision tree is a tree like collection of nodes intended to create a decision on values affiliation to a class. Each node represents a splitting rule for one specific Attribute. For classification this rule separates values belonging to different classes. The building of new nodes for both operators is repeated until the stopping criteria are met.

III. DESIGN

A. *Optimize Selection*

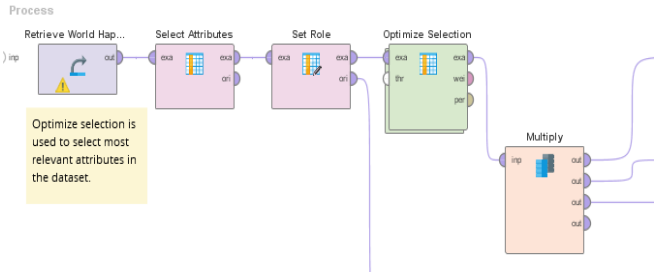

Figure 1: Retrieve World Happiness dataset with Optimize Selection.

Firstly, we retrieve World Happiness dataset and connect the output with select attributes. In select attributes parameter, we choose subset for attribute filter type. We only select eight from thirteen attributes since another five attributes seems to be not important for our goal. The attributes selected is Country, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual.

The output of select attributes operator is connect with set role operator. In set role parameter, we choose Country for attribute name and label as a target role. For set role output, we connect with two operator which is optimize selection and multiply operator which will connect with Random Forest and Decision Tree.

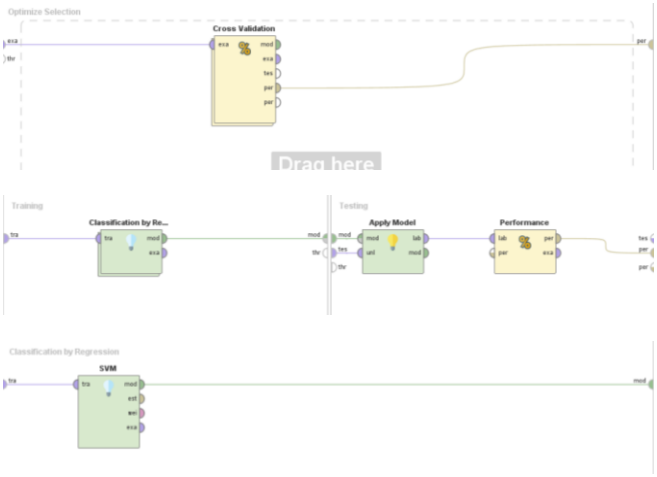

Figure 2: Subset of Optimize Selection.

The subset for optimize selection operator is cross validation operator. We connect the input from World Happiness dataset. Then, the performance connected with the output. The subset for cross validation operator is divide by two which training part and the other one is testing part. For training part, we used SVM which is subset of classification by regression. The output model connected with testing part which is apply model as it used to applying training model. The output of apply model is connect with performance operator. For the parameter on this part, we don't change anything since it already fix.

B. *Random Forest and Decision Tree*

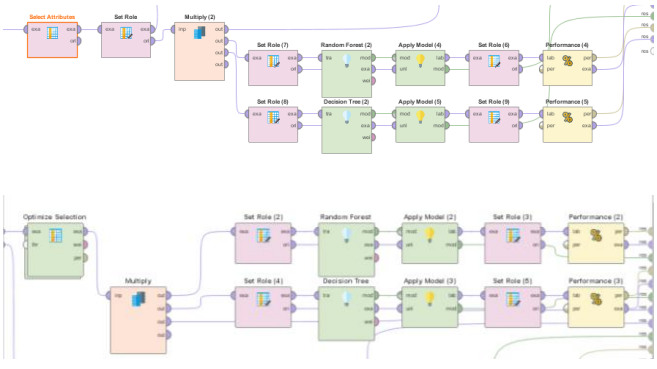

Figure 3: Retrieve World Happiness dataset and Optimize Selection with Random Forest and Decision Tree.

The result we get from selected attributes and optimize selection will connect to multiply operator since the result need to be used for both random forest and decision tree.

C. Correlation Matrix and Aggregate

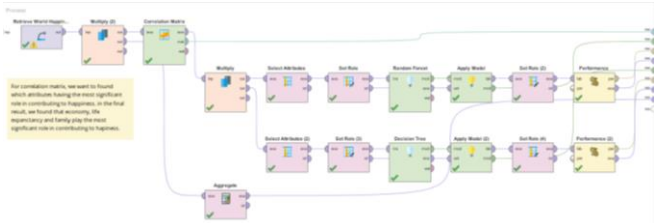

Figure 4: Retrieve World Happiness dataset, Correlation Matrix and Aggregation with Random Forest and Decision Tree.

The parameter of correlation matrix is subset for attribute filter type. We choose eight attributes which is Happiness Score, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual.

For aggregate part, the parameter in aggregation attributes we choose only tree which is Economy (GDP per Capita), Family and Health (Life Expectancy). All aggregation functions for aggregation attributes are average. The group by attributes is Country.

The result we get from correlation matrix will be connect with both random forest and decision tree by using multiply.

IV. RESULT

A. Optimize Selection

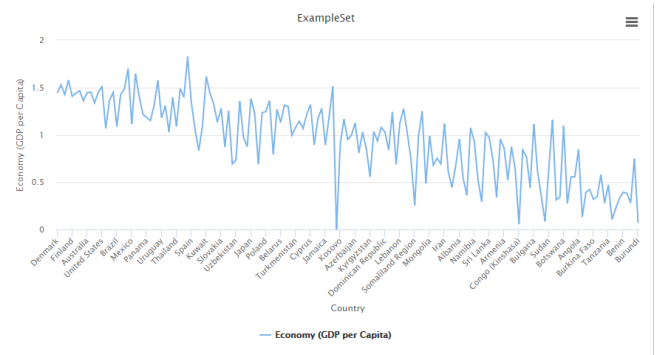

Figure 5: The result of Optimize Selection.

The purpose we used optimize selection is to select the most relevant attributes in World Happiness dataset. The result we get after we run the process is only Economy (GDP per Capita). It shows that the highest country happiness effected by the chosen attribute is Dubai.

B. Random Forest and Decision Tree Performance

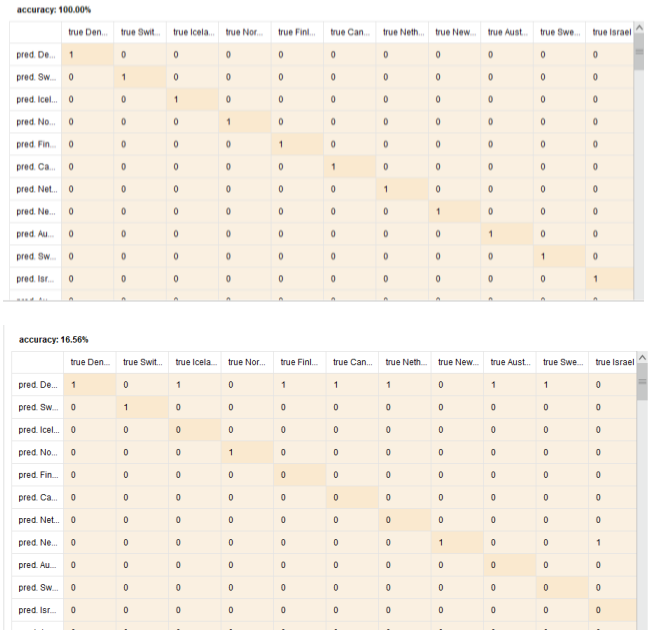

Figure 6: Result of Random Forest and Decision Tree Performance from Optimize Selection

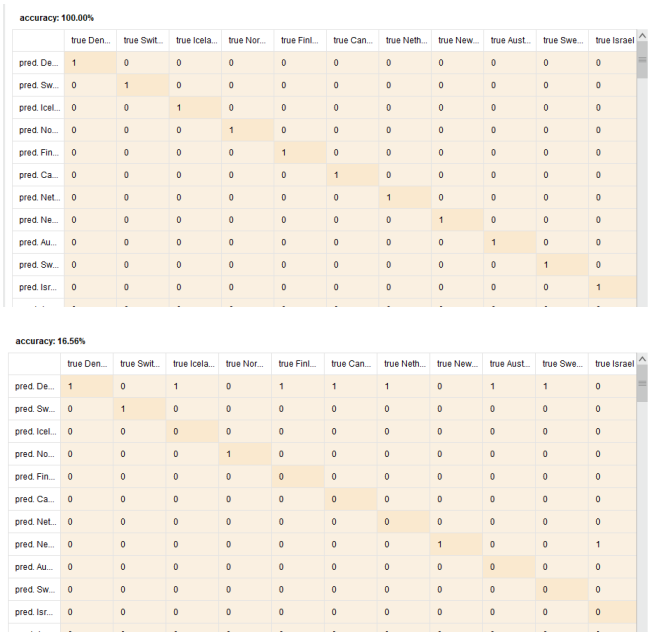

Figure 7: Result of Random Forest and Decision Tree Performance from Select Attributes.

The result of performance we get from random forest is 100.00% while from decision tree is 16.56%. Between this both operators, we know that random forest produce the best result than decision tree.

C. Correlation Matrix

First Attribute	Second Attribute	Correlation
Happiness Score	Economy (GDP per Capita)	0.790
Happiness Score	Family	0.739
Happiness Score	Health (Life Expectancy)	0.765
Happiness Score	Freedom	0.567
Happiness Score	Trust (Government Corruption)	0.402
Happiness Score	Generosity	0.157
Happiness Score	Dystopia Residual	0.544

Attribut...	Happine...	Econom...	Family	Health (...	Freedom	Trust (G...	Genero...	Dystopi...
Happine...	1	0.790	0.739	0.765	0.567	0.402	0.157	0.544
Econom...	0.790	1	0.670	0.837	0.362	0.294	-0.026	0.069
Family	0.739	0.670	1	0.588	0.450	0.214	0.090	0.120
Health (L...	0.765	0.837	0.588	1	0.341	0.250	0.076	0.101
Freedom	0.567	0.362	0.450	0.341	1	0.502	0.362	0.092
Trust (G...	0.402	0.294	0.214	0.250	0.502	1	0.306	-0.003
Generosity	0.157	-0.026	0.090	0.076	0.362	0.306	1	-0.133
Dystopia...	0.544	0.069	0.120	0.101	0.092	-0.003	-0.133	1

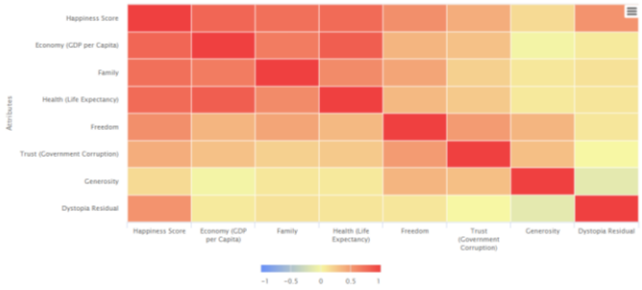

Figure 8: The result of Correlation Matrix.

The result shows that there are three attributes having the most significant role related to Happiness Score which is Economy (GDP per Capita), Family and Health (Life Expectancy) and Family.

D. Aggregate

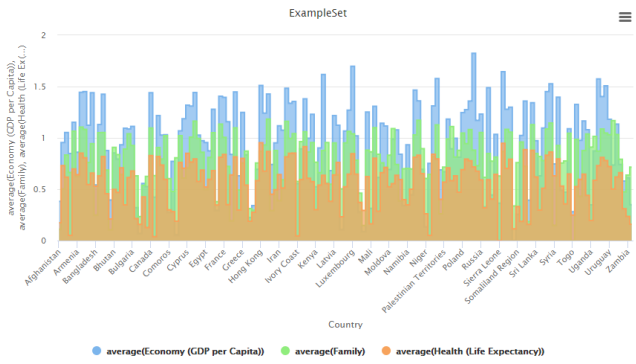

Figure 9: Result of average in three attributes which is Economy (GDP per Capita), Family and Health (Life Expectancy) and Family.

The result shows that the highest country happiness caused by Economy (GDP per Capita) is Dubai with 1.82427 average and the lowest country is Somalia. For Family attribute with the highest average which is 1.18326 is Iceland and the lowest country is Togo. While the highest country effected by Health (Life Expectancy) is Hong Kong with 0.95277 average and the lowest country is Sierra Leone. All the lowest country has 0 average effected by Economy (GDP per Capita), Family and Health (Life Expectancy).

V. CONCLUSION

The proposed World Happiness Ranking Prediction is to determine which countries or regions rank the highest in overall happiness, to estimate the six factors contribute highest happiness and to observe any country experience a significant increase or decrease in happiness. According to the result we get, the most attribute contribute in happiness of a country is Economy (GDP per Capita) and the highest country affected by the attribute is Dubai.

ACKNOWLEDGEMENT

We officially acknowledge that the Gallup World Poll as the data provider for the World Happiness Raking Prediction. In accordance to that, the development of this project is to fulfill the requirement to finish the Machine Learning Mini Project.

REFERRENCES

[1] Javad Zabihi (2018, May). Happiness 2017 (Visualization + Prediction) from https://www.kaggle.com/javadzabihi/happiness-2017-visualization-prediction

[2] Aggregate from Rapidminer Documentation, https://docs.rapidminer.com/9.1/studio/operators/blending/table/grouping/aggregate.html

[3] Correlation Matrix from Rapidminer Documentation, https://docs.rapidminer.com/9.1/studio/operators/modeling/correlations/correlation_matrix.html

[4] Decision Tree from Rapidminer Documentation, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

[5] Random Forest from Rapidminer Documentation, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html

[6] Optimize Selection from Rapidminer Documentation, https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection.html