World Happines Data Analysis

Hamza Nadeem. Author, Letterkenny Institute of technology

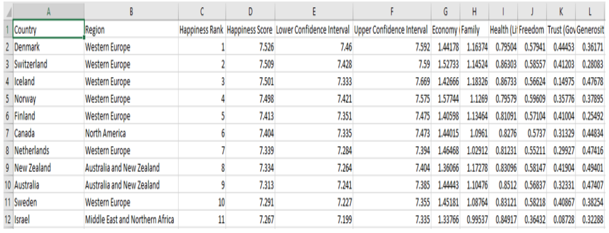
***Abstract*— The aim of this analysis is to decide which variables to live the happier life are more important. It helps people and countries to concentrate on the more important factors in terms of achieving a greater degree of satisfaction. We will also incorporate many machine learning algorithms to predict the rate and compare what algorithms work best for that individual dataset.**

***Index Terms*— spark, visualization, big data, analysis, data set**

I. Introduction

THE dataset We chose is of happiness 2016, which we get from Kaggle's datasets. The dataset offers the rating of happiness and happiness in 155 different countries based on seven indicators, including ”family”, ”life expectancy”, ”economy”, ”generosity”, ”trust in government”, ”freedom”. We decided this because we think the goal is to explore the scientific basis of assessing subjective well-being and understanding. we decided to focus on the first World Happiness study that accompanies the United Nations high-level meeting on harmony and individual well-being. The International Day of Happiness of the United Nations is published annually. This data is important because our health is influenced by a sense of purpose. Though both are important, positive emotions matter more to us than the absence of negative emotions. Our collective goal is to find out what major factors caused country rankings or scores to change between the reports for 2015 and 2016, and why countries change each year. we also want to know which ones or regions rate the high in general happiness and every one of the six key factors to happiness and have noticed a huge increasing or decreasing in happiness in any state. We aim at our target factor, that is happiness value, more specifically, and this information will enable us to evaluate it from thousands of individuals around the globe and explore ways to make people happy.

II. Data understanding



	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	Region	Happiness Rank	Happiness Score	Lower Confidence Interval	Upper Confidence Interval	Economy (	Family	Health (Li	Freedom	Trust (Gov	Generosit
2	Denmark	Western Europe	1	7.526	7.46	7.592	1.44178	1.16374	0.79504	0.57941	0.44453	0.36171
3	Switzerland	Western Europe	2	7.509	7.428	7.59	1.52733	1.14524	0.86303	0.58557	0.41203	0.28083
4	Iceland	Western Europe	3	7.501	7.333	7.669	1.42666	1.18326	0.86733	0.56624	0.14975	0.47678
5	Norway	Western Europe	4	7.498	7.421	7.575	1.57744	1.1269	0.79579	0.59609	0.35776	0.37895
6	Finland	Western Europe	5	7.413	7.351	7.475	1.40598	1.13464	0.81091	0.57104	0.41004	0.25492
7	Canada	North America	6	7.404	7.335	7.473	1.44015	1.0961	0.8276	0.5737	0.31329	0.44834
8	Netherlands	Western Europe	7	7.339	7.284	7.394	1.46468	1.02912	0.81231	0.55211	0.29927	0.47416
9	New Zealand	Australia and New Zealand	8	7.334	7.264	7.404	1.36066	1.17278	0.83096	0.58147	0.41904	0.49401
10	Australia	Australia and New Zealand	9	7.313	7.241	7.385	1.44443	1.10476	0.8512	0.56837	0.32331	0.47407
11	Sweden	Western Europe	10	7.291	7.227	7.355	1.45181	1.08764	0.83121	0.58218	0.40867	0.38254
12	Israel	Middle East and Northern Africa	11	7.267	7.199	7.335	1.33766	0.99537	0.84917	0.36432	0.08728	0.32288

With the 2015 and 2016 World Happiness Survey, we have various corporate concerns to address, as well as questions. We have to do a thorough analysis of our data sets to explain our output. We are provided two datasets to actually solve our scientific problem and respond to our queries. Our both datasets comprise rankings and happiness scores which use data from both the ”Gallup World Poll”. The data sets are focused on the important questions raised by life assessments in the survey. Questions are focused on the Cantril scale, asking participants to think about a ladder with the best life possible. This demands that participants rate their lives from nil to ten. Ten is the strongest possible life and the worst is nil.

III. Data Pre processing

A. Data Cleaning

We began to upload our files to jupyter to prepare this data set. First of all, we wanted to understand the lack of our results. We passed all files and found that we didn't have any missing values. First, we matched the two files we had to make sure the data were compatible. Drilling down into the data set, we noticed that little modification was necessary, given that the data was readable, full, consistent and accurate. It was very pristine

B. Data integration and reduction

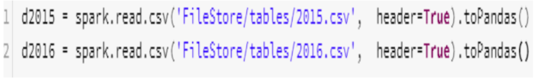
By searching for some of the more insightful variables we kept planning this dataset. We also conducted multiple experiments to classify the most insightful variables and to compare them to see if the more informative variables were related. First, we tried to see if we could delete or merge any of the variables. We then discussed how some different variables could theoretically be combined to build more reliable and perhaps insightful variables. We found, however, that because of the classification of each variable, we could not delete or merge variables. We addressed the class attribute as well.

C. Data Transformation

We found it easy to understand and reduced the time we spent pre-processing, provided that our data was already usable for each year and cleaned. Our data then began to be analyzed.

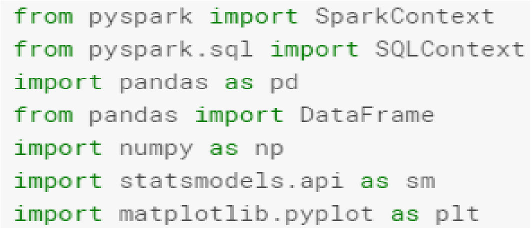
IV. Descriptive Analytics

Happiness is an essential element of ours living. We can examine what makes a person really happy through factors. Descriptive analytics are included in this. In describing what happened with our data, descriptive analytics are relevant. Given our two datasets, the libraries we used for visualizing our data were pandas, numpy, matplotlib. It gave us the best performance in our dataset exploration. it gave us an insight into our data set and an understanding of the events of 2015 and 2016.



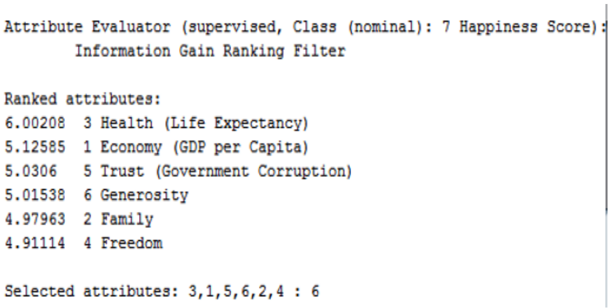
```
d2015 = spark.read.csv('FileStore/tables/2015.csv',  header=True).toPandas()
d2016 = spark.read.csv('FileStore/tables/2016.csv',  header=True).toPandas()
```

We have clearly found three significant variables that contribute to worldwide happiness. Such factors include ”health”, ”economy” and ”trust in government”. We modified the class variable to the happiness (numeric) when we were using python and spark.



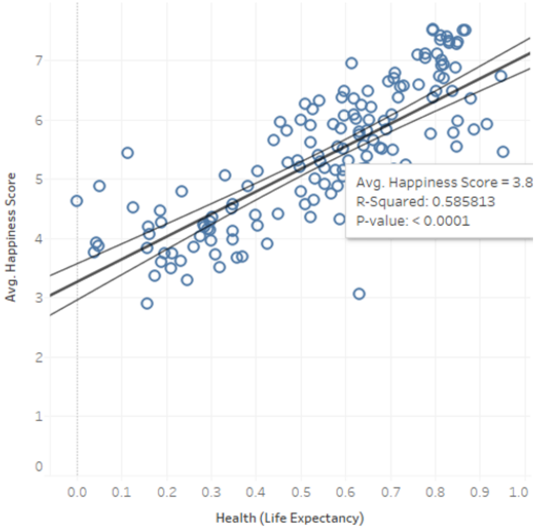
```
from pyspark import SparkContext
from pyspark.sql import SQLContext
import pandas as pd
from pandas import DataFrame
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

We then used this data set to examine the chosen attribute evaluator. We decided to carry out the assessment of the information gain, because the key variables influencing happiness scores across countries really interested us. We found several valuable information while operating this evaluator. Health was the variable with the most information. It has been classified first and shows we are very happy about health. That makes more sense logically. Economy is the 2nd most helpful variable. Trust was the third important info variable we found. Generosity is the fourth most insightful feature. Community was the fifth most helpful. Lastly, equality is the last quite insightful variable. From data we get that these observations align with the overall acquisition of knowledge. we must focus on and evaluate the three most important variables. The most insightful three variables are ”health”, ”economy” and ”trust”.

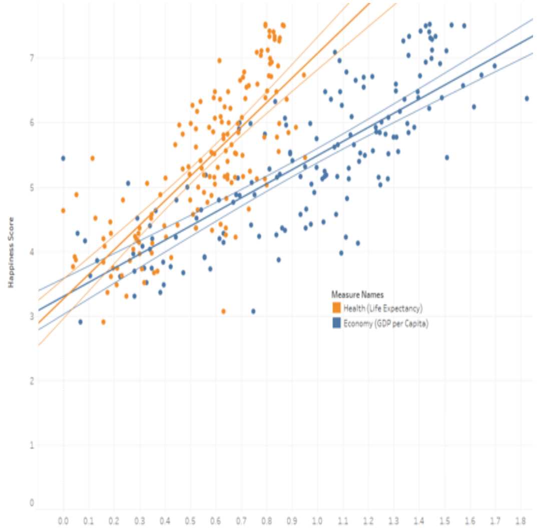


```
Attribute Evaluator (supervised, Class (nominal): 7 Happiness Score):
        Information Gain Ranking Filter

Ranked attributes:
6.00208  3 Health (Life Expectancy)
5.12585  1 Economy (GDP per Capita)
5.0306   5 Trust (Government Corruption)
5.01538  6 Generosity
4.97963  2 Family
4.91114  4 Freedom

Selected attributes: 3,1,5,6,2,4 : 6
```

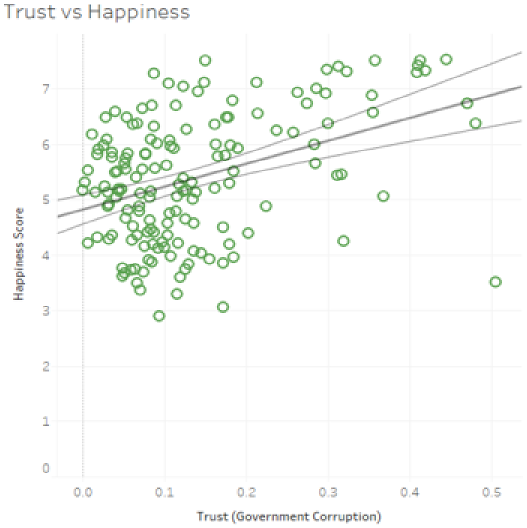
As we’re seeing below, the most informative variable of this data set is health. We can then use that information to verify if our target variable was associated to our most insightful variable. We graphed health and happiness result to assess this. A positive correlation can be seen from this. The overall happiness score increases as health tends to increase. It gives us a glimpse into why people in some countries are happy. Healthcare is an important factor. This is however not the only component which has a strong relationship with the happiness levels of a country.



From Python also we find that a relationship exists between our goal (happiness) variable and our next very insightful (economy) variable. We have also got an actual association between the GDP and the happiness Score. The overall happiness increases with an overall economic score and GDP in a region. This is rational because if you have a high GDP, then you will actually be satisfied. You’re less concerned and less doubtful. If your per capita GDP is low, your gladness is probably low. Vice versa. You are going to have problems to get through and have less satisfaction because of it.

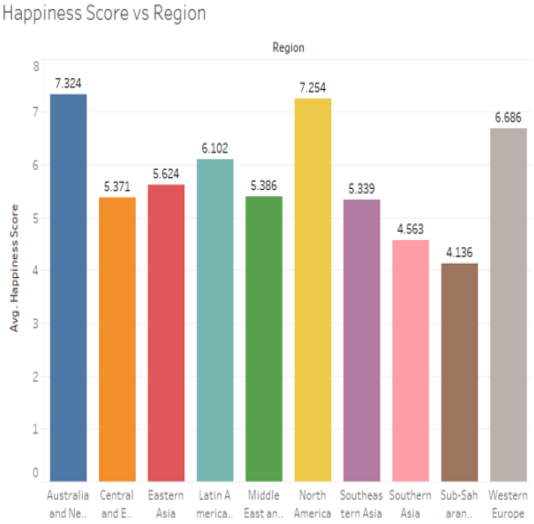


Trust (political corruption) has been my third most informative variable. The higher this attribute, the more the country has confidence in its governance. This information was then visualized in python and a trend was found to get an accurate presentation of the graph. We also found positive correlation between confidence and happiness after introducing the trend line. Thus more people have faith in their government, the better it is. It makes total sense because the less happy you are if you are upset and have little confidence in your government. These data were however gathered before the election of Donald Trump to the Unites states May not be recent for happiness ranking. The relationship between increases and decreases was also identified between 2015 and 2016.



When we analyzed the data, we noticed that certain levels of happiness decreased, increased or stayed the same throughout 2015 and 2016. We expected to analyze what caused a country to be more or less happy. By analysing data between 2015 as well as 2016, we can analyze the main contributing factors to increasing and decreasing the happiness number. While some outliers are present, most genetic differences are the same. Their three most insightful variables also increased when the satisfaction of a country improved.

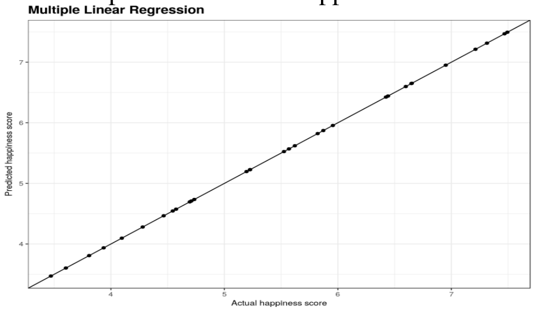
The following diagram shows the connection between our target happiness and region. The graph shows that Australia has the greatest average level of happiness, followed by North America. The four leading regions all illustrate all the key factors that contribute to joy of country: "caring", "freedom", "generosity", "honesty", "health", "income" and "good governance". For estimates, the grades can be re-ordered from year to year, even tiny changes.



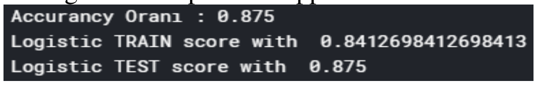
V. Predicting happiness using ML models

We'll use several machine learn algorithms in this segment to predict happiness. Next, our dataset set should be divided into training and test set. Our conditional variable is happiness, and it is "family", "economy", "life expectancy", "trust", "freedom", "generosity", and "dystopia residual".

The overview demonstrates that all independent variables play a major role, and R squared measured is 1! Since we have mentioned, a linear correlation among dependent and separate factors is clear. Also, the amount of the independent variables corresponds to the dependent variable, which is the incentive factor. This is the reason why an adjusted R is exactly identical to 1. As a test, Multiple Linear Regression is predicted to predict 100 percent accurate happiness score

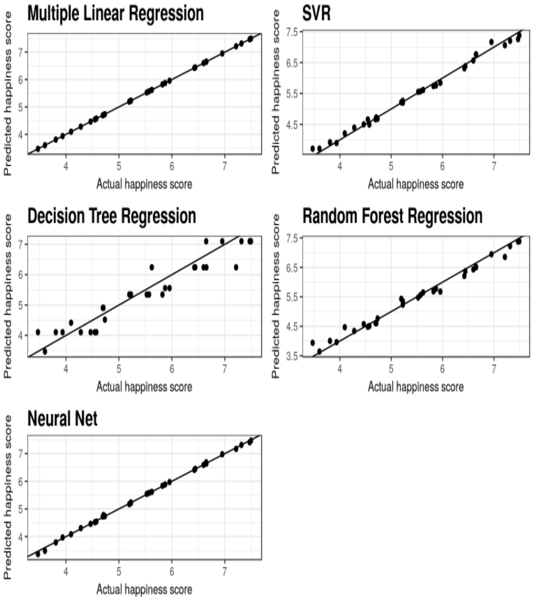


The logistic regression model forecast fairly high accuracy happiness scores. "Multiple linear regression" and neural networks have done a very good job and have projected about the same result. In terms of predictability, SVR and Random Forest stood second. The Decision Tree finally could be the worst algorithm to predict happiness.



```
Accurancy Oranı : 0.875
Logistic TRAIN score with  0.8412698412698413
Logistic TEST score with  0.875
```

following result of different graphical result of machine learning algorithm can easily demonstrate that which model has a high accuracy rate of which one has a worst.



VI. Conclusion

Our findings varied from our initial ideas. However, we understood the errors in our initial hypothesis after the various modeling techniques were applied. Our the use of such a information gain ratio and our careful revision and combination of certain attributes such as happiness ranks made us carefully avoid adjusting and other negative influences to obtain a clear reading of the relationship among the most informative variables. We contrasted our most insightful variables to each other in order to identify the competence coefficient between variables. This helped us to accurately assess our performance. Our methods of experimenting with various models both with and without boosts have enabled us to better understand how we can assess, implement and deploy our results later. Because this dataset was based on an abstract notion, we deemed all of the different business areas covered by our dataset and chose two which ultimately affected everyone else: capitalism and govt. We also learned that many businesses have developed internationally into new areas. Globalization brings new resources and opportunities and with our data set analysis we think we can assist any organization in expanding its market. Our data set assessment helped us to understand the connection between the key happiness variables and how they correlate. This knowledge of the happiness factors would allow a business to determine where to place new stores. We figured that the foundation for the world happiness study for opening a new location had the chance to increase ROI. A business can choose where to open a new place on the basis of a country's overall happiness level by means of thorough analysis of our research. For instance, if an international retail shoe company were to open its first shop but could not choose which country, the model we created could be used for determining a new place based on the economic standing, population and health of the country.

References

[1] Rai, D. (2020). Logistic Regression in Spark ML. [online] Medium. Available at: https://medium.com/@dhiraj.p.rai/logistic-regression-in-spark-ml-8a95b5f5434c [Accessed 23 Feb. 2020].

[2] Kaggle.com. (2020). Machine Learning with Apache Spark. [online] Available at: https://www.kaggle.com/lpdataninja/machine-learning-with-apache-spark [Accessed 23 Feb. 2020].

[3] GitHub. (2020). lp-dataninja/SparkML. [online] Available at: https://github.com/lp-dataninja/SparkML/blob/master/kaggle-titanic-pyspark.ipynb [Accessed 23 Feb. 2020].

[4] Zabihi, J. (2020). Happiness 2017 (Visualization + Prediction). [online] Kaggle.com. Available at: https://www.kaggle.com/javadzabihi/happiness-2017-visualization-prediction [Accessed 23 Feb. 2020].

[5] Kaggle.com. (2020). Analysis of World Happiness. [online] Available at: https://www.kaggle.com/kralmachine/analysis-of-world-happiness [Accessed 23 Feb. 2020].

[6] Mishra, R. (2020). Learning plotly while analysing happiness score. [online] Kaggle.com. Available at: https://www.kaggle.com/dataraj/learning-plotly-while-analysing-happiness-score**Finding-from-All-Analysis** [Accessed 23 Feb. 2020].

[7] Stack Overflow. (2020). how to do multiple target linear regression in Spark MLLib?. [online] Available at: https://stackoverflow.com/questions/44170572/how-to-do-multiple-target-linear-regression-in-spark-mllib [Accessed 23 Feb. 2020].

[8] GitHub. (2020). susanli2016/PySpark-and-MLlib. [online] Available at: https://github.com/susanli2016/PySpark-and-MLlib/blob/master/Linearregressionhouse.ipynb [Accessed 23 Feb. 2020].

[9] Spark.apache.org. (2020). MLlib: Main Guide - Spark 2.4.5 Documentation. [online] Available at: https://spark.apache.org/docs/latest/ml-guide.html [Accessed 23 Feb. 2020].

[10] Databricks. (2020). Machine Learning Library (MLlib) – Databricks. [online] Available at: https://databricks.com/glossary/what-is-machine-learning-library [Accessed 23 Feb. 2020].

[11] Dayananda, S. (2020). Spark MLlib — Machine Learning In Apache Spark — Spark Tutorial — Edureka. [online] Edureka. Available at: https://www.edureka.co/blog/spark-mllib/ [Accessed 23 Feb. 2020].

Hamza N. Author Hamza Nadeem is currently doing his MSc. in big data analytics and artificial intelligence from letterkenny institute of technology. he done his bachelors in computer science. his major research interest in the field of artificial intelligence specially in the domain of deep learning and computer vision. before his masters he is working as a unity 3D developer and developed augmented reality applications for the company.

Under the supervision of Dr. Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, CO. Donegal Project link: https://github.com/hamzaexe/World-Happiness-analytics