Clustering world happiness and predicting it by Random Forests

Emmanuel Avila Orozco A01704617
Tecnológico de Monterrey Campus Querétaro
November 14th, 2022

Abstract

This document explains the usage of different tools in data analysis such as Random Forest for linear regression and clustering in order to separate and analyze data for predictions and interpretations.

Introduction

Happiness can be very subjective into different perspectives, but there's always a way to interpret and quantify data by doing analysis reports from organizations that take different measures using different factors such as the gross income from an entire country or government trust.

In that way, according to the World Happiness Report, the rankings of national happiness are based on a Cantril ladder (fig. 1) survey, which consist on taking nationally representative samples of various countries and being asked to think about the best possible life they could imagine, and then ask them by their actual qualification of their life perception, then the report correlates the life evaluation results with different life factors. (WHR | The World Happiness Report, 2022).

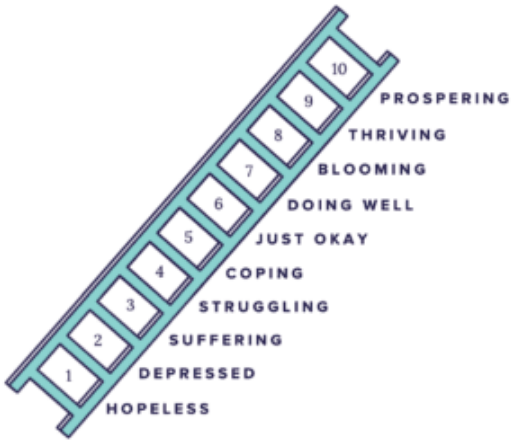

Fig. 1. Cantil Ladder example.

The main purpose of this project is to analyze and see how much weight some life factors have upon others to allow the user to predict how happy on a scale from 0 to 7 based on their life factors is. The first approach to use the dataset is to apply a PCA analysis and then try a decision tree for regression analysis and compare the performance with a random forest.

Finally a comparison between K-mean vs DBSCAN clustering is made in order to see how well the information can be categorized for future implementations.

Happiness and corruption dataset

According to the owner of the dataset (Elias Turk, 2022), this dataset

represents the main factors that can be the reason why some countries can struggle with prosperity and economic growth. Being happy can represent a benefit to all citizens because happier workers, happier results.

This dataset contains about 13 columns that describe the happiness of different countries from a lot of perspectives, such as freedom, corruption perception index, country, generosity, gdp per capita and so on.

On figure 2, it is displayed the different instances presented on the dataset, that further on will be edited in order to discriminate between them to see which ones can be useful to make predictions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 792 entries, 0 to 791
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Country            792 non-null    object
 1   happiness_score    792 non-null    float64
 2   gdp_per_capita     792 non-null    float64
 3   family             792 non-null    float64
 4   health             792 non-null    float64
 5   freedom            792 non-null    float64
 6   generosity         792 non-null    float64
 7   government_trust   792 non-null    float64
 8   dystopia_residual  792 non-null    float64
 9   continent          792 non-null    object
 10  Year               792 non-null    int64
 11  social_support     792 non-null    float64
 12  cpi_score          792 non-null    int64
dtypes: float64(9), int64(2), object(2)
memory usage: 80.6+ KB
```

Fig. 2. General information about the dataset.

A very good thing about this dataset is that there´s no null information about any class, and that makes it easier to work with it.

Visualizing data

As a first approach to this dataset, it was considered important to see how variables can affect the happiness score and understand in a way to make connections with previous knowledge and assumptions that help us to understand what is being displayed.

On figure 3, a good example of how different countries have a variation in happiness score is shown. This helps us to see what media and cultural information teaches us that may be wrong but also to see maybe the countries we look forward to as an example of prosperity can be not so well perceived by its inhabitants.

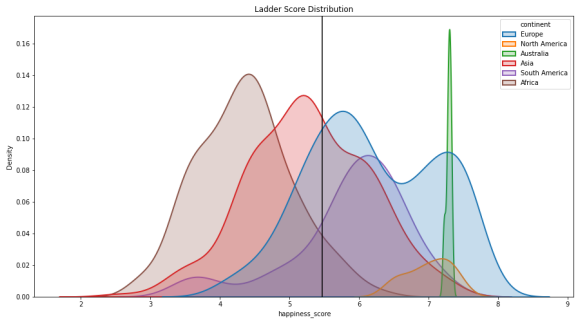

Fig. 3. Different happiness scores from continents.

Then, focusing more on the usage of the data, a matrix of Spearman correlation was made in order to see which attributes can help us to make better predictions because of the weight in the final score. As seen on figure 4, just by looking at the Spearman matrix it was easy to select some characteristics over others.

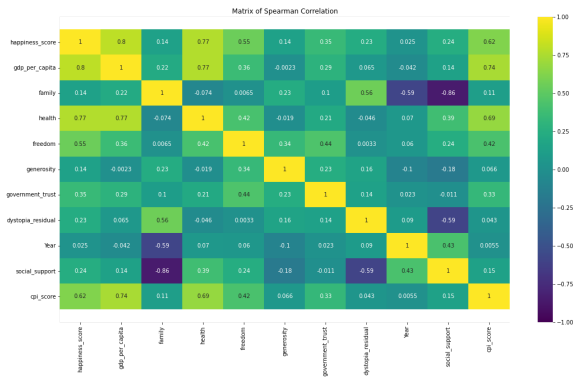

Fig. 4. Matrix of Spearman Correlation.

Looking for a model to predict data

According to Geeksforgeeks (2022), a decision tree is one of the most popular tool for classification and prediction because of its characteristics like the flowchart structure where each node denotes a type of test on an attribute and each branch represents an outcome of the test and each leaf node holds a class label.

On the other hand, as Tony Yiu (2019) says, random forest as the name implies, consists of a large number of decision trees that use the rule of the majority´s knowledge, making a better tool for prediction and classification than decision trees alone, in figure 5, a basic scheme of a random forest is displayed.

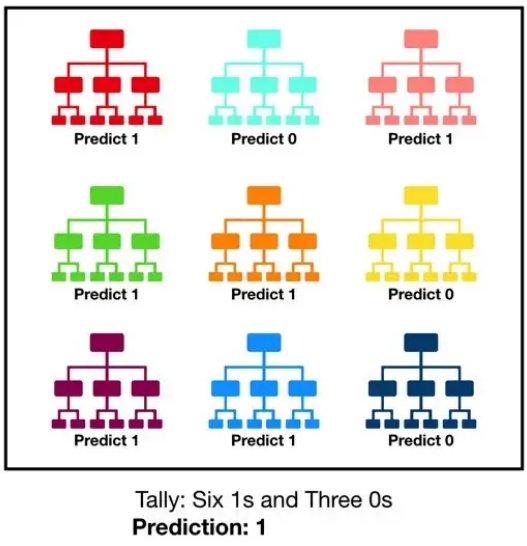

Fig. 5. Basic scheme of a random forest.

According to Jaadi (2022), Principal Component Analysis (PCA) is a dimensionality-reduction method very useful for large datasets and by transforming a large set of variables into a smaller one that preserves most of the important information, in other words: reduce the number of variables of a data set, while preserving as much information as possible, maybe some accuracy is lost in the progress but simplicity is gained in return for that.

Looking at the definition of GeeksforGeeks (2022) of clustering, the main idea is that this type of unsupervised learning method tries to draw references from datasets that consist of input data without any label response. Clustering is the task of dividing the population of data points into a number of groups such that data points in the same groups are more similar to other data points.

Proposed model

As a first approach for this data set was to try a decision tree for regression in order to see how well it performs even though the main focus of this type of algorithm is not for continuous data, but after cleaning the data and running a test to see their performance, the results can be seen in the image below.

```
regressor = DecisionTreeRegressor(random_state = 0)
regressor.fit(X_train, y_train)
y_Tree_pred = regressor.predict(X_test)
r2_score(y_test,y_Tree_pred)
```

```
0.7953774635044523
```

Fig. 6. Performance for decision tree.

then random forest

Then, a random forest was implemented to see if that very very large tree can be optimized and even to get better results for our prediction. One very important thing to remember is that even though the results may come good, it is the responsibility of the data analyst to discern if it is a good idea to take that result and use it as accurate.

```
RFregress = RandomForestRegressor()
RFregress.fit(X_train, y_train)
y_test_pred = RFregress.predict(X_test)
print('Testing accuracy on all features: %.3f' % r2_score(y_test, y_test_pred))

Testing accuracy on all features: 0.919
```

Fig. 7. Performance for random forest.

After getting a good sense of this method, it was decided to perform a prediction based on the random forest as a regression tool. For that, a set of variables need to be introduced by the user and then work with that information to make the desired prediction, in the image below, an example of a run is displayed.

```
Predict your actual happiness based on metrics from your country
How much does GDP contributes to your happiness? (0-3): 3
How much does Life Expectancy contributes to your happiness? (0-1.15): 1.15
How much does Corruption Perception Index contributes to your happiness? (0-91): 91
How much does your freedom perception contributes to your happiness? (0-0.73): 0.73
How much does Government Trust perception contributes to your happiness? (0-0.6): 0.6
How much does Social Support perception contributes to your happiness? (0-1.65): 1.65
Based on the saddest country, how much happier you think ypu are? (0-3.6): 3.6

 In a scale from 0-7, your happiness score is:  6.962

 Do you want to make another prediction? (Y/N): n
```

Fig. 8. Prediction made by the random forest.

After several tests, the results were nearly good enough to start making another analysis of the outcome and see some differences between the results. In order to do that, a clustering between two important factors were made, by pure election of variables as seen in the Spearman matrix, the happiness score and the GDP per capita were the most correlated variables, below there's a figure plotting some points in order to make analysis by K means clustering.

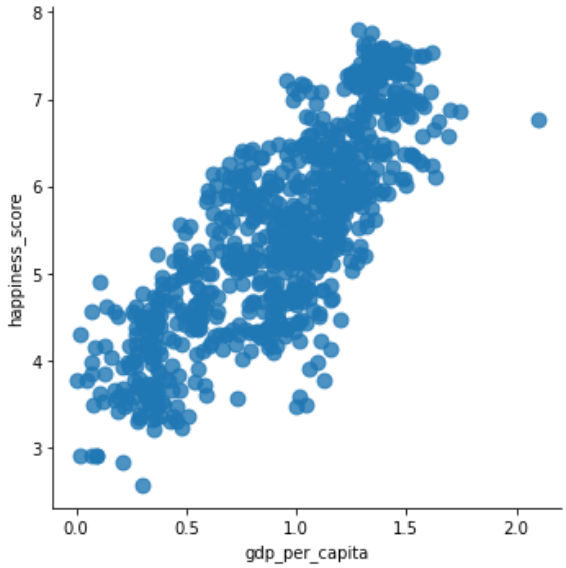

Fig. 9. Plot of instances between GDP per capita and happiness score.

In order to make a consistent analysis, the data taken before from the user was added into the data frame but now for an analysis that locates the type of country the user may be in. It's important to denote that there´s only two attributes working on this clustering, and not the entire data set, mainly because of a dimensionality problem that may be talked about further in this document.

A K-means clustering was applied using the new data from the user in order to classify the country the person belongs to, the best appropriate division of clustering was just 2 groups. In figure 10 an example of the clustering is displayed.

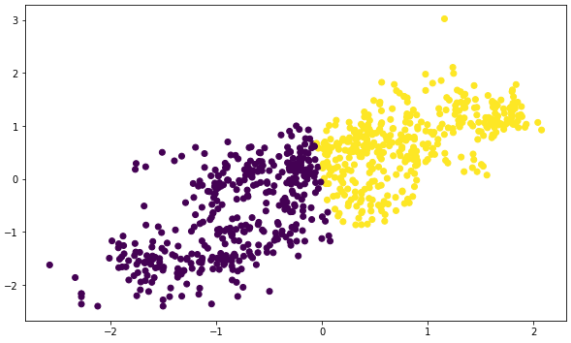

Fig. 10. Plot of clustering grouping using user data.

As it is seen, two different groups were made that can more or less have an impact on the happiness score that the user has. After a run of the code by using the user inputs, here in the image below are some results.

Fig. 11. Results after clustering with user data.

As seen above, the results from running the program with user data coincides with the predictions before, meaning that the happiness score may be accurate.

As mentioned before, when dealing with huge data sets it's difficult to see many dimensions at the same time, so PCA analysis is taken upon the data. First, with a new data set that doesn't take into account some variables that have no effect on the results, then a PCA between those 7 columns is made to see how they correlates.

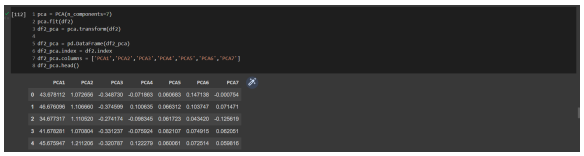

Fig. 12. PCA dataframe.

After that, data was scaled and severe plots were made to see how the clusters may interact with each other, the most easier to group and make clusters of it, was the one that compared PCA2 vs PCA3, with an arrangement of 3 clusters, that fitted very well as can be appreciated below.

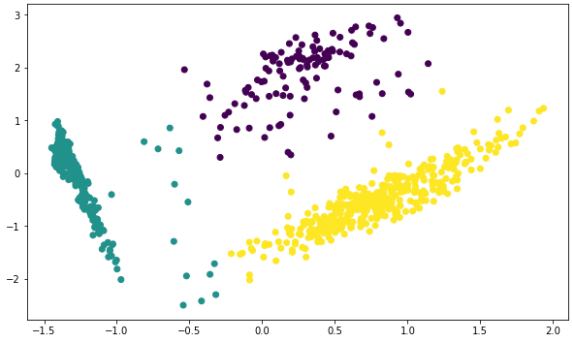

Fig. 12. Scatter plot of PCA2 vs PCA3.

With that preview, the same method was desired to complete with the entire dataframe of PCA to see how it can react so then a technique of clustering can be made, first it was using K-means and then with DBSCAN to see what can be achieved.

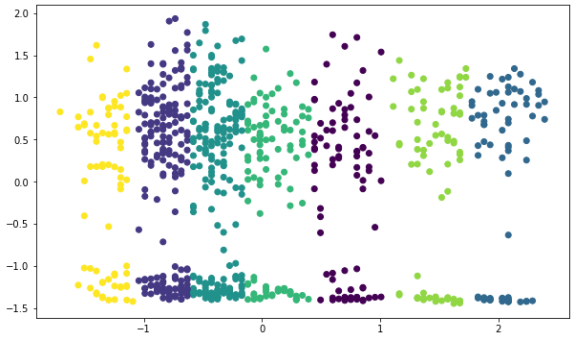

Fig. 13. Scatter plot of PCA data frame using K-means.

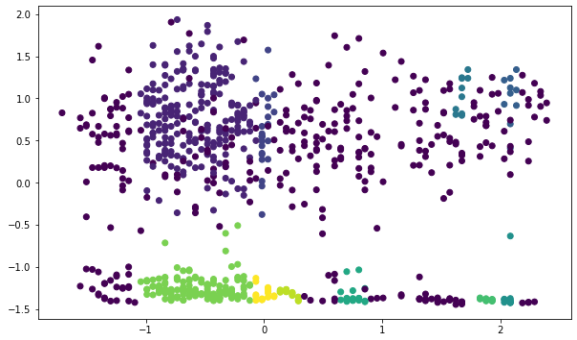

Fig. 14. Scatter plot of PCA data frame using DBSCAN.

Conclusions

After all the analysis and visualization of the data, different tools were used successfully to predict data from the user inputs, PCA analysis and clustering that data provide good information and understanding of the project as a whole.

For me, as a mechatronics engineer, this whole course was a challenge. I´ve never encountered this type of content of machine learning and data analysis to perform such interesting algorithms using probability and software engineering. Nonetheless I wish I had the basis to understand the classes better from the professor because it was joyful to see how someone so good in this field can teach the basics.

Bibliography

Home | The World Happiness Report. (2022, 12 septiembre). https://worldhappiness.report/

GeeksforGeeks. (2022, 4 octubre). *Decision Tree*. https://www.geeksforgeeks.org/decision-tree/

Yiu, T. (2019, 12 junio). *Understanding Random Forest*. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Jaadi, Z. (2021, 1 abril). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis