Classifying countries based on demographic variables with decision trees

Hugo Pérez, A01273106, Tecnológico de Monterrey Campus Querétaro

Abstract - This document presents the implementation and explanation of a classification employing random forests algorithms, as well as a benchmark between of this algorithm under different conditions

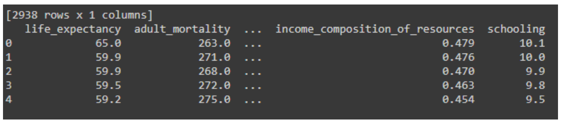
I. INTRODUCTION

Globalization is an essential part of our lives. Now more than ever it easier to have access to resources and communicate from all around the world. This reflects for many countries as a positive boost to its development. Despite being able to cooperate and collaborate globally, some countries have still issues and are considered behind in terms of development, thus giving room for the classification terms "developed countries" - those that have achieved advanced technological infrastructure, industrialization, widespread infrastructure and a general high standard of living; and "developing countries" - that are yet to reach that prosperity level.

II. DATASET

The dataset used was obtained from Kaggle(https://www.kaggle.com/kumarajarshi/life-expectancy-who) and is a compilation of data collected from the World Health organization and the United Nations website.

It consists of 2938 record of different countries across several years, with 20 attributes containing demographic characteristics, income composition and mortality rates. Figure 1 shows an example of the data with some of its attributes.



```
[2938 rows x 1 columns]
   life_expectancy  adult_mortality  ...  income_composition_of_resources  schooling
0             65.0            263.0  ...                            0.479       10.1
1             59.9            271.0  ...                            0.476       10.0
2             59.9            268.0  ...                            0.470        9.9
3             59.5            272.0  ...                            0.463        9.8
4             59.2            275.0  ...                            0.454        9.5
```

Figure 1. Example of the dataset

Full list of attributes:
country,
year,
status,
life_expectancy,
adult_mortality,
infant_deaths,
alcohol,
percentage_expenditure,
hepatitis_b,
measles,
bmi,
under-five_deaths,
polio,
total_expenditure,
diphtheria,
hiv/aids,
gdp,
population,
thinness_1-19_years,
thinness_5-9_years,
income_composition_of_resources,
schooling

III. APPROACH

The purpose of this implementation is to come with a model that can classify countries in the developed or developing categories based on the demographic data of each country.

This classification task can be performed with a decision tree, which will be the first step. then a random forest will be used to see how the accuracy can improve. A Principal Component Analysis will be performed to determine if it is possible to achieve the same (or a better) accuracy with a less complex model.

To keep most of the data for the classification, missing values will be filled using the mean strategy. The name of the country and the year of the sample will be dropped since are not relevant for the classification.

IV. PCA

To determine if it is useful to reduce dimensions, the first thing we do is plot the correlation between all the variables shown in Figure 2. The numbers we are interested in are those colored lime/yellow, for having a direct correlation and purple, for having an inverse correlation.

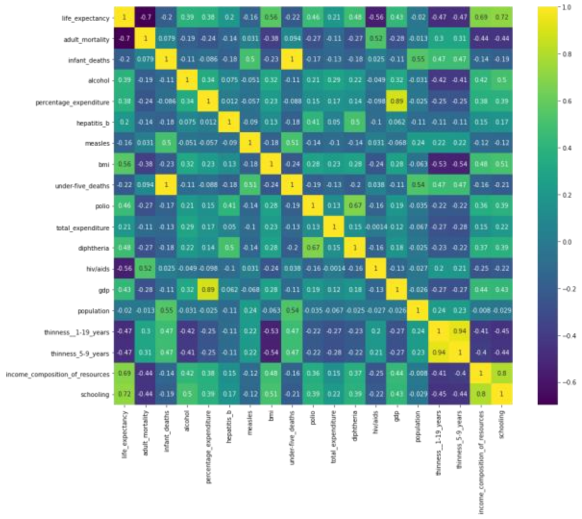


Figure 2. Correlation map

Most of the data variables are close to a 0 correlation which is not interesting for this analysis. There are a couple pairs like life expectancy with hiv/aids, adult mortality, income composition of resources and schooling that can help reduce a couple of dimensions in the model.

Plotting this pairs of variables together confirms the correlation between them.

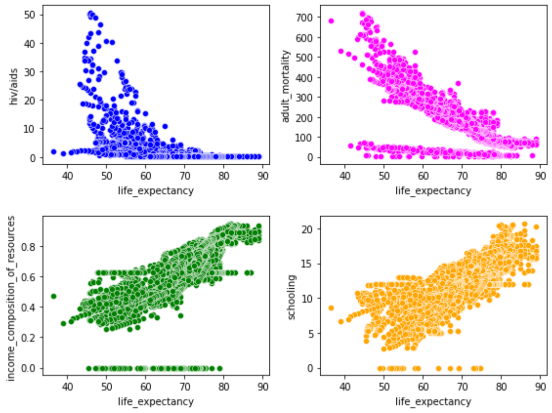


Figure 3. Correlations. Blue: life expectancy and hiv/aids. Magenta: life expectancy and adult mortality. Green: life expectancy and income composition of resources. Orange: life expectancy and schooling.

Running a the PCA and plotting the total variance explain for each component shows that 90% - 95% of the variance in the data can be explained with a number of components between 10 and 13, from there not much information is gain in the rest of the components.

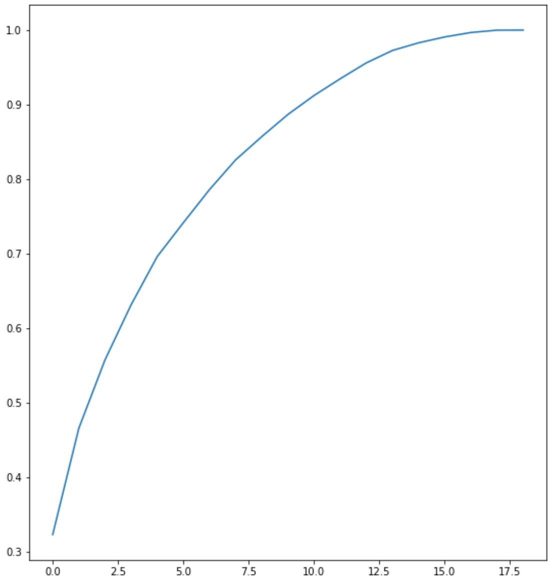


Figure 4. Percentage of variance explained by each component

Knowing this the goal is to reduce the 20 dimensions that the original dataset had to a 12-component model.

V. DECISION TREE

Using a single decision tree with 5 layers of depth and the whole attributes can get an accuracy of 93%. Adding more depth can increase the accuracy up to 96%, but since the goal is to optimize the model, we will use 5 depth as reference.

Training the model with the PCA components and using a random forest with 12 trees of depth 5 we can increase the accuracy to 94% while being able to reduce the dimensions from 20 to 12.

Figure 5 shows the tree obtained in the single tree run and gives us an idea of how the structure of the trees on the random forest is.

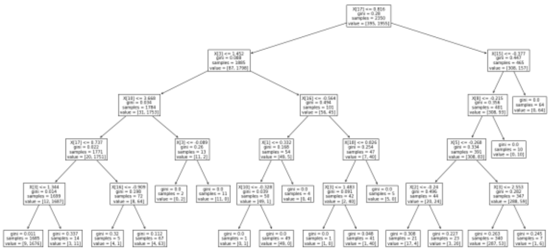


Figure 5. Decision tree example

VI. CONCLUSION

Based on the PCA, some of the variables are correlated and so we can optimize the model to work with 12 components while keeping the original accuracy.

Increasing the number of layers in each tree or the number of trees in the model could increase the accuracy, but at the cost of a more resource demanding model.

We can confirm that the demographic variables presented to explain life expectancy also define the characteristics to differentiate between a developed country and a developing one. This can be useful for governments and/or organizations like the WHO or the UN to determine where the efforts for development should focus on each country.

VII. REFERENCES

[1] "*Life Expectancy (WHO)*". Kaggle.com. (2017). [Online]. Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who

[2] *"PCA and the #TidyTuesday best hip hop songs ever".* Julia Silge. (2020). [Online]. Available: https://juliasilge.com/blog/best-hip-hop/

[3] "*Implementing PCA in Python with Scikit-Learn".* Usman Malik. (N/D). [Online]. Available: https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/