Using Classification and Clustering to Model Economic Development Status

By: Lloyd Page and Colin Schultz

Introduction

Countries are generally divided into 2 categories regarding economic development, “developed” and “developing”. Conventionally, these divisions are done somewhat arbitrarily where the latter classification is generally assigned to a country with a low industrial base and low Human Development Index. However, as noted by the United Nations itself, “there is no established convention for the designation of "developed" and "developing" countries or areas in the United Nations system”.

Related Work

There has been significant amounts of work in medical fields and economics among others where countries have been separated into developed and developing countries. For example, we have “Differences in prevalence, awareness, treatment and control of hypertension between developing and developed countries” by Pereira, Marta, et al. in epidemiology, which uses this distinction to find such differences. Or consider “ICT and economic growth – Comparing developing, emerging and developed countries” by Niebel, Thomas in economics. In the latter case, there is the addition of emerging economies, which are economies that are transitioning between the two other categories. In this paper, we will be keeping emerging economies as developing economies to simplify things. Niebel’s paper winds up showing that on ICT and economic growth, the distinction proves irrelevant, whereas Pereira’s paper finds the distinction to be significant. This paper hopes to add to the already large body of literature on the relevance of the distinction with regards to certain variables, as well as possibly answering the question of whether having exactly 2 or even 3 groupings (if one

separates emerging economies) is indeed the correct approach when separating countries into groups.

Methodology

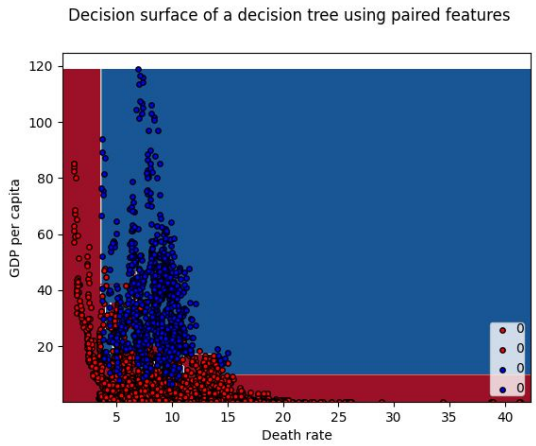
The goal of this paper is to use classification and clustering methods to test how arbitrary these designations are, and if there is a better way to group countries. To do the first, we will run a classification algorithm on data from the world bank and look at the accuracy of the algorithm where countries are classified based on whether or not the International Monetary Fund designated them as an “advanced economy” at the time the data was taken, and also the complexity of the decision boundary and the margin of the data points from said boundary. We will then run a clustering algorithm on the same data sets to determine if 2 groupings are appropriate.

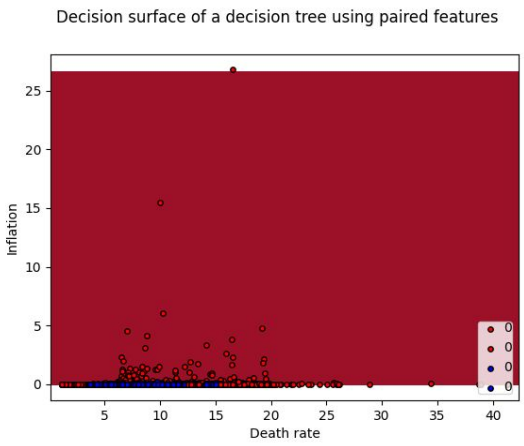
The Data comes from the world bank and the IMF. We chose 5 variables to analyze from the world bank based on the amount of data each variable contained and the variables being unique and rate-based. This resulted in the selection of Death rate, Inflation rate, Unemployment rate, Urban Population (%), and GDP per capita. Despite selecting indicators with more data than most, there were still many missing values, and even some missing countries (most notably, North Korea). Unfortunately, we lack the background and technical knowledge to be able to replace these missing values with appropriate values, and instead had to opt in to dropping them. This resulted in most of our data being concentrated into a 28 year stretch from 1991 to 2018, and the data set being much more likely to contain a developed country than what one would expect it to contain naturally. However, this shouldn’t particularly impact the analysis, as we still have 4875 data points, of which 871 represent developed countries and 4004 represent developing countries. All of these values should be large enough to do a proper analysis. We opted to use the International Monetary Fund’s status of “Advanced

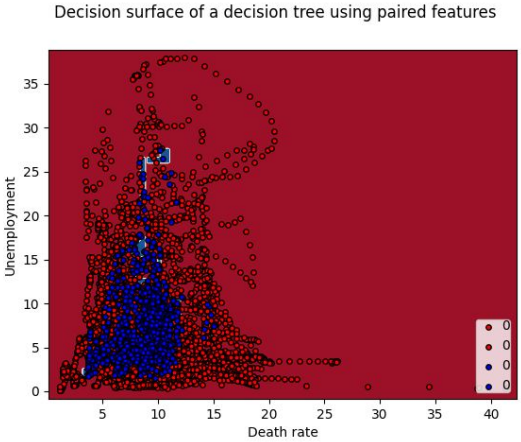
Economy” to represent a developed economy, as it had easily accessible historical data, making it easier to handle countries that have transitioned from developing to developed during the time frame of available data, such as the Asian Tigers and some countries in Eastern Europe.

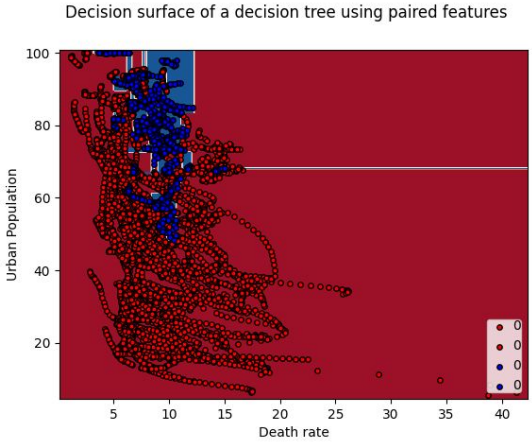
Classification

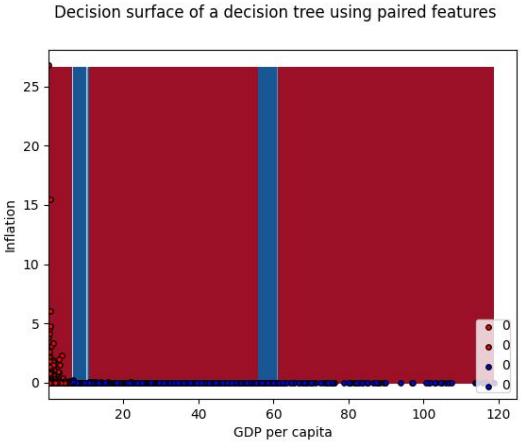
For Classification we opted to use a decision tree classifier. This was selected because decision tree classifiers are generally easier to understand than most classifiers in their decision making process, and during preliminary trials, it had the highest accuracy %, averaging around 97%, of the clustering methods tried, which were KNeighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net, AdaBoost, and QDA. Surprisingly, the more complex and advanced classification techniques failed to outperform more basic approaches with respect to the dataset. It is worth noting that the final decision tree is not particularly simple, having a depth of 13, which made it so complex that in-lining it with readable values is impossible, instead, we opted to showcase its selection method through a graphical representation. It was trained on 60% of the dataset, randomly selected, and having a 0.9717948717948718 chance of correctly identifying test data, with a 0.9813432835820896 chance on developing countries and 0.9269005847953217 chance on developed countries.

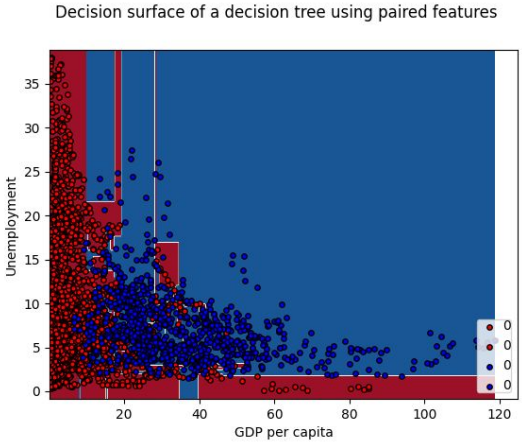


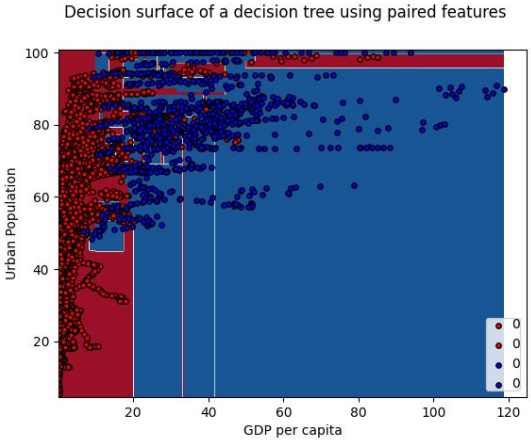


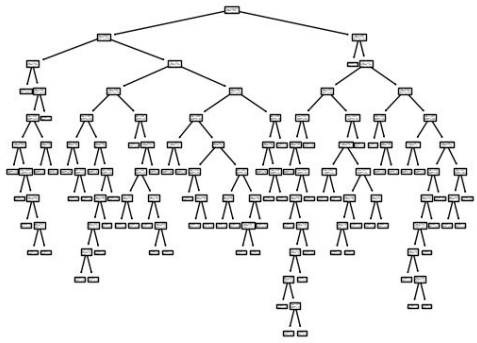

Decision surface of a decision tree using paired features
Unemployment
Death rate
0
0
0
0

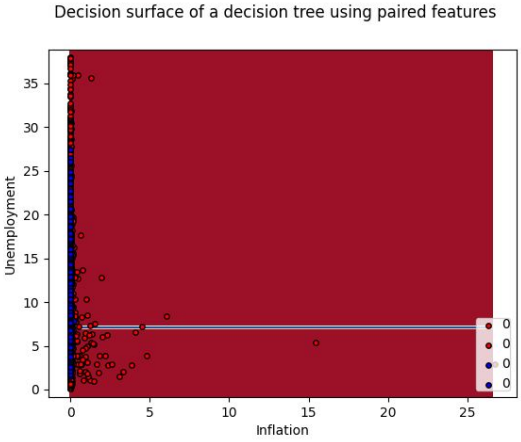

Decision surface of a decision tree using paired features
Urban Population
Death rate
0
0
0
0

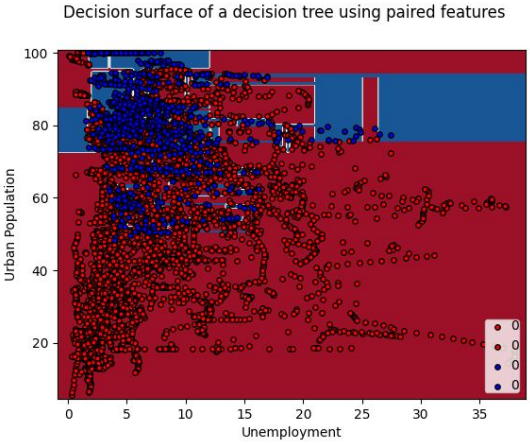

Decision surface of a decision tree using paired features
Inflation
GDP per capita
0
0
0
0

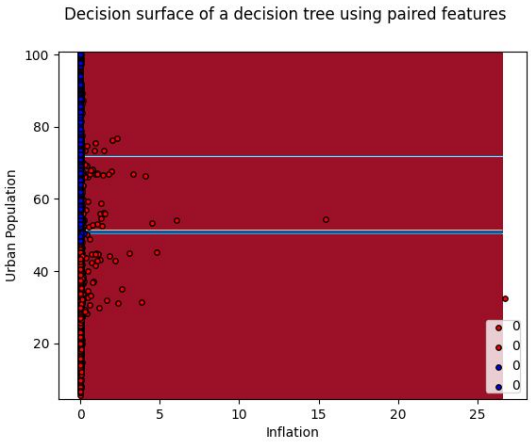

Decision surface of a decision tree using paired features
Unemployment
GDP per capita
0
0
0
0


Decision surface of a decision tree using paired features
Urban Population
GDP per capita
0
0
0
0




Decision surface of a decision tree using paired features
Unemployment
Inflation
0
0
0
0


Decision surface of a decision tree using paired features
Urban Population
Unemployment
0
0
0
0

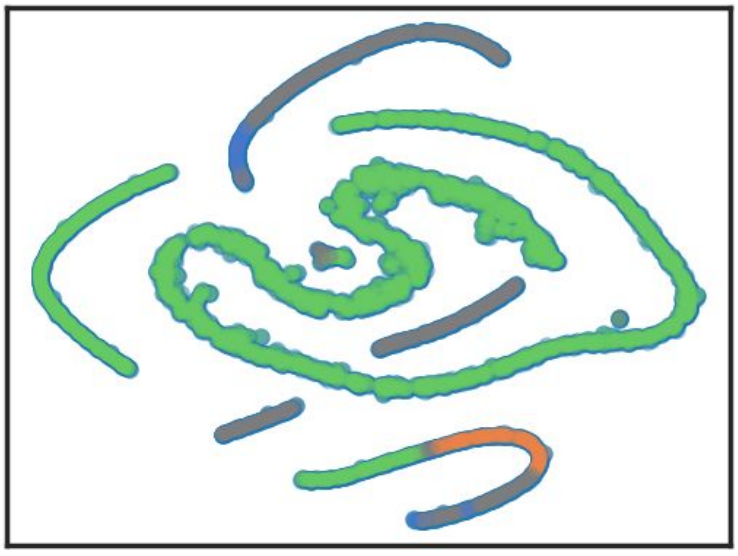


In the graphs above, the red areas are for developing countries, the blue for developed. We also opted to include the tree structure of the classifier. The smaller rectangles designate leaf nodes. It is worth noting that the inflation and GDP per capita values were normalized for the visualization, but not the classification algorithm. This was done due to a lack of processing power. There is some noticeable clustering in the 2-D graphs shown above. It is worth noting that the algorithm was more successful on developing countries than developed, which is surprising, as you would expect the opposite as countries that reach the statistical plateaus to get recognized would probably have to wait before getting recognized. This certainly could have arisen from the general lack of developed countries making it harder for the algorithm to properly adjust the model for them. In general, looking at the graphs, it appears that the margin of data points from the decision boundary was very small, which could also have contributed to less accuracy in developed countries, as the small margins probably resulted in the decision boundaries being better tuned for developing countries instead of developed countries.

Clustering

For clustering we used a Hierarchical Density Based Scan. This algorithm starts off by creating a reachability matrix for all the nodes. It then creates clusters based upon a minimum spanning tree it creates between all nodes in the data. To measure the distance between all of the nodes, we used simple Euclidean distances as all our data is on a percentage scale from 0 to 100. It then creates a hierarchical structure by sorting the nodes by the highest distance first. It then starts to create clusters from there. The way it does this is by creating a connected line from the most reachable point where it can get a certain amount of nodes within a cluster. This is a predetermined value which we choose as 100 nodes for a minimum cluster. Any nodes that do not fall within one of these clusters are determined as noise or unreadable. It finally assigns each node to a cluster and returns the list of them.

To visualize our data, we used manifold learning to transform our 5th dimensional data onto a 2nd dimensional plane. It takes a random projection of the data that was run through an algorithm to try and preserve distances. We specifically use t-distributed Stochastic Neighbor Embedding as our dataset is small enough for the algorithm to run at a quick pace.



The clustering algorithm is very good at clustering developing countries together. This is shown as the green cluster above. The orange and blue clusters are developed countries that were clustered. We were only able to correctly cluster 125 developed countries out of the 871 total data points, however, we could properly cluster 3591 of the 4004 developing countries. We were left with 962 points of data that fell off and were thought of as noise as seen by the grey data points above. We had a very low error rate of only incorrectly identifying 197 data points. We believe that it is much harder to cluster a developed country as this could be due to the data being too vast or our sample size for them was much smaller so the algorithm had a harder time finding the cluster.

Conclusion

We had several unexpected results, specifically with regards to developed countries, which certainly could be a by-product of the relative lack of data we had on developed countries. The other alternative suggested by the results of the clustering algorithm is that there may be 2 different categories of developed countries. However, it is difficult to say that with any degree of certainty, given how few developed country data points were placed at all, and the fact that fewer than 1 in 8 such points were placed. The classification algorithm results seem to suggest that the division between the 2 groups as currently defined holds up well, although the lack of complete separability suggests some misclassification over the years. The sheer complexity of the decision tree and lack of a large margin for many points from the decision boundary may suggest that the separation is artificial, but the fact that it held up well against testing data makes that less likely. Possible extensions to this paper are including more variables to see if the results to still hold as dimensionality and variables considered increase, increasing the data considered by filling in the missing data using background knowledge instead of throwing it out as we did, and doing statistical tests to see if there truly are 3 clusters in the data.

References

Campello R.J.G.B., Moulavi D., Sander J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg

Niebel, T. ICT and economic growth – Comparing developing, emerging and developed countries. In: World Development, Volume 104, April 2018, Pages 197-211

Pereira, M. Lunet, N. Azevedo, A. Barros, H. (2009) Differences in prevalence, awareness, treatment and control of hypertension between developing and developed countries. In: Journal of Hypertension: May 2009 - Volume 27 - Issue 5 - p 963-975

IMF. (1998) IMF Advanced Economies List. World Economic Outlook, May 1998, p. 134

IMF. (2001) World Economic Outlook, April 2001, p. 157

IMF. (2007) World Economic Outlook, April 2007, p. 204

IMF. (2008) World Economic Outlook, April 2008, p. 236

IMF. (2009) World Economic Outlook, April 2009, p. 184

IMF. (2011) World Economic Outlook, April 2011, p. 172

IMF. (2012) World Economic Outlook, October 2012, p. 180

IMF. (2014) World Economic Outlook, April 2014, p. 160

IMF. (2015) World Economic Outlook, April 2015, p. 48

IMF. (2016) IMF Advanced Economies List. World Economic Outlook, April 2016, p. 148

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

United Nations (2020). Standard country or area codes for statistical use (M49) retrieved on April 20th, 2020 from https://unstats.un.org/unsd/methodology/m49/

World Bank. (2020) GDP per capita (current US$) retrieved on April 20th, 2020 from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Bank. (2020) Inflation, GDP deflator (annual %) retrieved on April 20th, 2020 from https://data.worldbank.org/indicator/NY.GDP.DEFL.KD.ZG?view=chart

World Bank. (2020) Unemployment, total (% of total labor force) (modeled ILO estimate), retrieved on April 20th, 2020 from https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=chart

World Bank. (2020) Death rate, crude (per 1,000 people), retrieved on April 20th, 2020 from https://data.worldbank.org/indicator/SP.DYN.CDRT.IN

World Bank. (2020) Urban population (% of total population), retrieved on April 20th, 2020 from https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?view=chart