Analysis of Countries Vulnerable to Climate Change

Jeff Mathew
Department of Computer Science
University of Warwick

Abstract– **Climate change is one of the foremost challenges facing mankind today. From gradual effects like global warming to more immediate threats such as flooding and wildfires, the impact of climate change cannot be ignored. Moreover, studies have shown that this impact is not distributed uniformly across the globe. This paper aims to study the different characteristics of countries across the world to try and identify whether any patterns exist that make some countries more vulnerable to climate change than others.**

1 INTRODUCTION

Climate change is a very hotly debated topic and rightly so. The decisions and policies of today have the potential to affect the lives of millions across the globe for decades to come. Numerous experts have suggested the world will soon be upon the point of no return - that is a point in time where the damage done is irreversible[1].

While any measures tackling climate change are undoubtedly a boon for the whole world, it is important to keep in mind that time is running out quickly for some countries compared to others. Unfortunately, some parts of the world are at greater risk to the effects of climate change than others. The analysis published by *Germanwatch* annually [2] ranks the countries of the world based on how vulnerable they are to climate change's more adverse immediate effects, primarily extreme weather events like floods and heatwaves. This ranking leverages information regarding economic losses as well as fatalities to assign a *Climate Risk Index (CRI)* to countries across the world. The countries at the top of this index are deemed to be more vulnerable to adverse effects of climate change.

The CRI is a reflection of the results of climate change. In other words, it is analysis done post extreme events to rank countries based on vulnerability. The aim of this paper is to look at different socioeconomic and geographic characteristics of these countries to determine whether there are any patterns that innately make some countries more vulnerable to climate change than others. Any such patterns would suggest that it is not just random chance that some countries are at greater risk.

2 DATA SOURCES

Two data sources have been used in this study and they're described below.

2.1 Climate Risk Index (CRI) Data

The CRI data is obtained from the analysis published by *Germanwatch* for the year 2021 [2]. This index is used to distinguish between countries that are considered to be high risk and those which are considered to be low risk.

2.2 Country Characteristics Data

This generic dataset was obtained from Kaggle and provides information such as GDP, mortality rates, literacy rates etc for each country in the world [3].

3 HYPOTHESIS

Although the effects and behaviour of climate change can be difficult to predict, the general hypothesis is that certain countries are more vulnerable for a reason. More specifically, we explore the two hypotheses below:-

a) Geographical location of a country highly influences its risk to disasters related to climate change. Some regions of the planet are more vulnerable than others and the countries occupying these regions are at higher risk.

b) Socioeconomic factors can also play a role in determining risk. Less developed countries may not have quality healthcare to support themselves during a disaster or due to being economically weaker might struggle more to recover from a climate related tragedy.

Condition	Class
CRI ≤ 50	High Risk
CRI > 50	Low Risk

Table 1. CRI Rank to Class Conversion

4 TOOLS USED

The software tools used to aid the analysis performed in this study are listed below:-

4.1 Jupyter Notebook

Some of the initial data cleaning and pre-processing (described in the next section) was done using Jupyter notebooks, especially with the help of the Pandas library.

4.2 Weka

Weka is a powerful data visualisation and modelling tool which can be used to train/evaluate different types of machine learning models and iterate quickly.

5 DATA CLEANING

a)The first task is to combine the information from the two data sources into a form such that machine learning techniques can be applied on it. The CRI data does not include every country in the world - the countries for which the authors could not evaluate all the required metrics for were discarded from the CRI rankings. In other words, the list of countries in the CRI data is a subset of those in the Kaggle countries dataset (which contains information for all countries). Hence, from the Kaggle dataset only records corresponding to countries present in the CRI data were kept. This resulted in a total of 175 countries for which all the required data is present.

b)The CRI data essentially provides a ranking. To perform some supervised classification tasks and evaluate the quality of clusters generated, this ranking is translated into two classes, namely *low-risk* and *high-risk*. These classes were defined based on the simple and intuitive logic described in Table 1.

c) *Normalization* - The different features in the Kaggle countries dataset such as population and square area of the country are in completely different scales. This would definitely affect the performance of the clustering algorithms used which employ Euclidean distance metrics to evaluate the similarity between different instances. Most classification techniques also benefit from having features restricted to a common range. Decision tress are an exception since the model simply decides a split value at each level and the scale of the value for the split is irrelevant to the result. The normalization was carried out with the *filters* option in Weka.

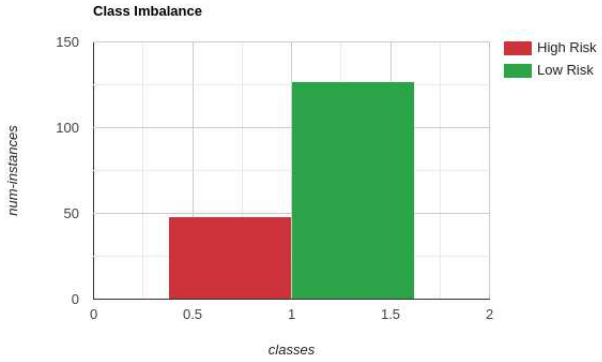

Fig. 1. Bar graph representing the class distribution across whole dataset

d) *Class Imbalance* - As mentioned earlier, a total of 175 countries are present in the final pre-processed dataset. Of this only countries in the top 50 of the CRI are classified as *high-risk* with the remaining being classified as *low-risk*. This means that there is a heavy class imbalance and the number of low-risk instances are more than twice the number of high-risk instances. A popular approach to mitigating this issue is to use a technique called SMOTE [4] which creates artificial samples similar to those already present in the dataset.

However, creating instances of 'fake' countries and using them to train ML models does not seem like a very intuitive approach as metrics such as GDP and mortality rates are dependent on a variety of factors. Since the focus here is primarily on the high-risk countries (the minority), it was decided to keep the original dataset as it is and instead weight the instances such that the minority class is given just as much importance as the high-risk class. The option of assigning a much larger weight to the minority class was considered but dropped considering the limited number of samples available which would just end up training models that fit the minority class perfectly. The instance weighting was again achieved using the *filters* option in Weka.

e) *Feature Selection* - As a very basic feature elimination step, the feature representing the names of the countries are dropped as they provide no valuable information and are just 175 unique string values. Later during the classification stage, a relatively more comprehensive *backward selection* approach is employed to identify the features most relevant to the high-risk class.

Please see Appendix-A for the list of all features included in the dataset for analysis after the data cleaning stage.

6 DATA VISUALIZATION

Some features from the dataset that represent geographical information and socioeconomic factors are visualized below to illustrate how the two classes are distributed for those

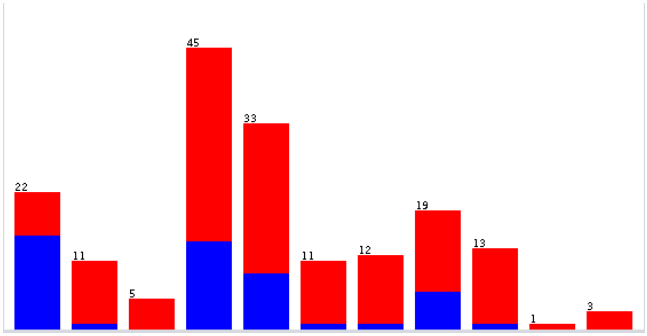

Fig. 2. Class distribution with respect to different regions

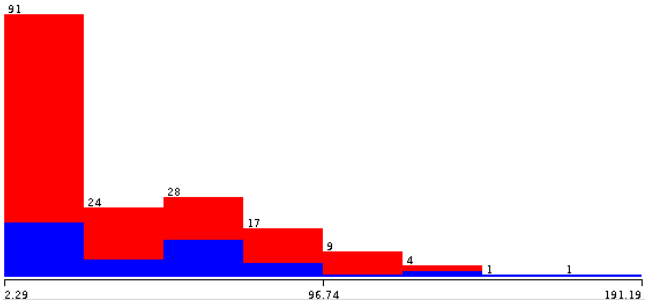

Fig. 3. Class distribution with respect to infant mortality rate

features. This is an initial analysis and a glimpse into what features maybe relevant to the analysis being done. In all the bar graphs discussed below in this section, the Blue bars represent the *high-risk* class and the Red bars represent the *low-risk* class.

6.1 Geographical Region

Some regions of the planet are potentially more vulnerable to climate disasters than others. It follows that the contries that belong to such regions are also probably at a higher risk of suffering from climate change's adverse effects. The dataset has grouped the countries into eleven different regions.

From Fig.2 it is clear that in certain regions the proportion of one class is much greater than the other (labels for regions have not been included in the plot to avoid clutter). For example, the bar on the extreme left corresponds to Asia and seems to suggest that countries in Asia are quite vulnerable. The second bar from the left represents Eastern Europe and here it looks like the opposite holds true and countries in the region are at less risk.

6.2 Infant Mortality Rate

Infant mortality rate can be connected to some degree with the development and healthcare facilities associated with a country. Less developed countries tend to have higher rates compared to more developed countries.

From Fig.3 it appears as though the high-risk instances make up a larger proportion of the bars as infant mortality rate increases. This maybe an indicator that less developed countries tend to be at higher risk.

6.3 Gross Domestic Product (GDP)

GDP is a strong indicator of the economic strength and resilience of a country. From Fig.4 it appears as though the trend observed with infant mortality rate is reversed. As GDP values increase, the bars tend to be made up mostly by low-risk instances. This suggests that economically strong countries are generally at lower risk of being adversely affected by climate change. This could also mean that in the event of a crisis well off countries have the means to recover more quickly compared to less developed nations.

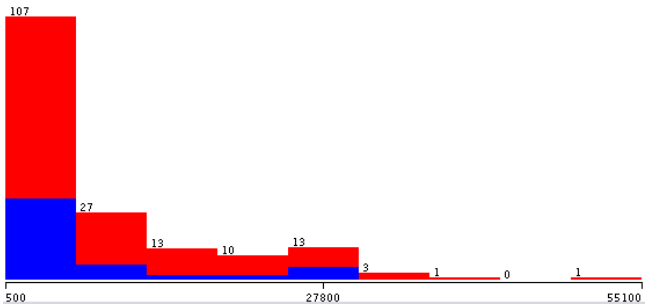

Fig. 4. Class distribution with respect to GDP

7 MACHINE LEARNING TECHNIQUES

This section describes the various ML approaches explored to try and draw insights from the data. To reiterate, this study is mainly focused on the high-risk (minority) class, with the objective of identifying any characteristics that tend to make a country more vulnerable to climate related disasters and thus climate change.

As mentioned in section 5, all the numeric attributes in the dataset have been normalized. Additionally, the instances have all been weighted such that the minority class is given just as much importance as the majority class. While experimenting, it was observed that both these pre-processing steps resulted in improved performance from the models.

The study aims to analyse the results of two ML techniques, namely clustering and classification. Both approaches are described in more detail below.

7.1 Clustering

Clustering is an unsupervised technique that will try to group the instances (countries) into clusters based on how similar they are to each other. Since it is already known that there are two classes in the dataset, a clustering algorithm should ideally identify two clear and distinct clusters.

The expectation is that the two clusters will result in a fairly good separation of instances belonging to both classes, in which case we can study the clusters to identify characteristics that distinguish high-risk instances from low-risk ones.

A number of clustering approaches were tried with the help of Weka: k-means, DBSCAN, Farthest first clustering and Expectation Maximization(EM). Both EM and DBSCAN figure out the optimal number of clusters while building the model. K-means and Farthest first algorithms require the number of clusters to be inputted before hand - two in our case here. In each case, the quality of the generated clusters were evaluated using the class labels as a reference. After multiple experiments and hyperparameter tuning it was found that the best performing clustering model was the Farthest First algorithm which achieved an accuracy score of about 66%.

However, considering the imbalanced nature of the dataset, simple accuracy score would not be a good metric to evaluate the results. Since the emphasis is on the minority class, the model picked as the best performing one also had the most correctly grouped high-risk instances among all clustering models. These results can be observed from Fig.5

Although there is coherence to the clusters, the two classes have not been separated perfectly by the clustering. Next, the clusters are visualized with respect to different attributes to identify any patterns and evaluate the hypotheses made in section 3. Cluster0 corresponds to the low-risk class and cluster1 the high-risk class.

Figures 6 through 9 illustrate some of the features that seem to be separated fairly well by the clusters. In Fig.6, it is observed that the low-risk cluster is actually densely packed where infant mortality rates are close to zero. On the other hand, the instances with higher mortality rates seem to belong to the high-risk cluster. Fig.7 shows that the GDP for almost all instances in the high-risk cluster are at the lower end of the range. Fig.8 highlights that there is quite a bit of separation based on the regions as well - the chances of finding a high-risk instance in some regions (such as the Baltics) is very low. The above three insights seem to align well with the initial hypotheses made. However, it is important to remember that the clusters have not separated the two classes perfectly, so there are quite a few samples in both clusters that are noisy.

Fig.9 is interesting in that this feature was not considered in our initial hypothesis. The number of phones per 1000 people can again be connected to some degree with the level of development and economic well being of a country. Once again, the data seems to be suggesting that the less developed countries are at higher risk of suffering from the effects of climate change.

7.2 Classification

Unsupervised methods, especially clustering is an intuitive way to analyse data. However, supervised approaches have the capability to nudge the model in the right direction during training as it is already aware what the right class for each instance is. In keeping with our overall objective, the aim with classification is to fit a good model to the data and then interpret the trained model to identify which features are weighted heavily by the model. Hopefully, this will provide some insight into the characteristics of both classes, especially the high-risk class. Since the instances have been weighted to account for class imbalance, a good classifier should be able to fit the minority class decently well.

```
 0  1  <-- assigned to cluster
31 17 | HIGH
98 29 | LOW

Cluster 0 <-- LOW
Cluster 1 <-- HIGH

Incorrectly clustered instances :       60.0     34.2857 %
```

Fig. 5. Farthest First clustering results

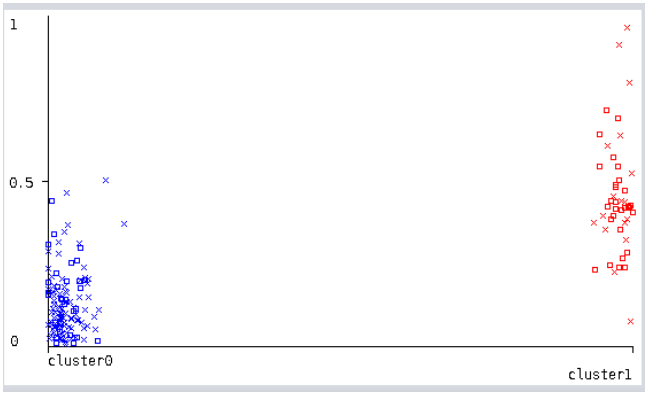

Fig. 6. Clusters with respect to Infant Mortality

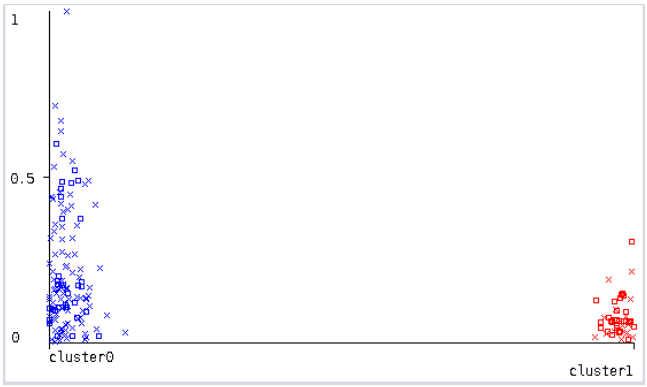

Fig. 7. Clusters with respect to GDP

Once again, a number of different approaches with different combinations of hyperparameters were experimented with using Weka. The techniques include Naive Bayes, SVM, Logistic Regression, Decision Trees and k-Nearest neighbours. In terms of evaluation metrics, there are a few more options available within Weka for classification tasks. Taking into account the heavy class imbalance, Area Under the Precision Recall Curve (AUPRC) has been chosen as the metric for comparison between models as this is proven

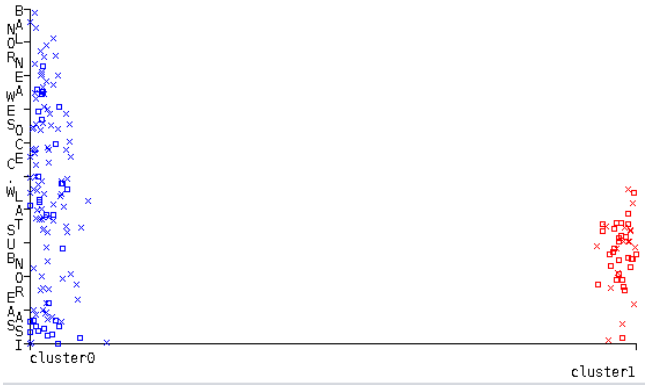

Fig. 8. Clusters with respect to different regions

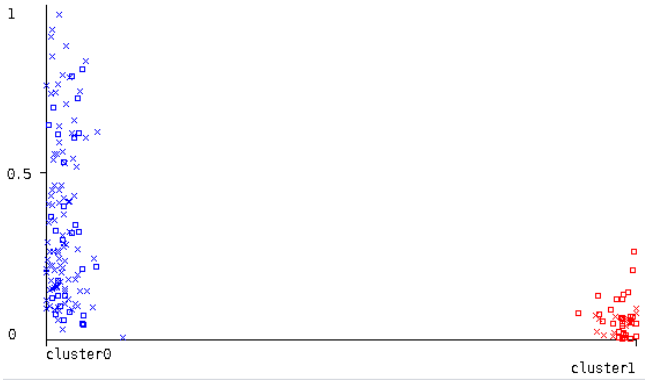

Fig. 9. Clusters with respect to phones per 1000 people

to be a somewhat robust metric for imbalanced problems. When the overall AUPRC scores are very similar between two models, the AUPRC score for the high-risk (minority) class is used to decide which model is better.

As mentioned before in section 5, a technique known as *backward selection* has been used to pick the suitable subset of features for the classification task. For each model, this was done by initially training a model using all the features in the dataset and then iteratively removing features one by one and observing how model performance changes. If performance improves or remains the same after the feature is removed, that feature is then dropped. If performance decreases when a feature is removed, the feature is kept to be used for training the final model (please see Appendix-A for a list of all features in the dataset). As a result of this process, the below five features were dropped from the dataset and this actually resulted in an improvement in model performance:-

a) Population
b) Arable (land)
c) Other
d) Industry
e) Service

Of the different approaches tried, Random Forest (an ensemble of decision trees approach) performed the best with an overall AUPRC score of 75.8% and a minority class

```
Correctly Classified Instances         122.8962               70.2264 %
Incorrectly Classified Instances        52.1038               29.7736 %
Kappa statistic                          0.4045
Mean absolute error                      0.4413
Root mean squared error                  0.4562
Relative absolute error                 88.2578 %
Root relative squared error             91.2385 %
Total Number of Instances              175

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.625    0.220     0.739     0.625     0.677     0.409     0.773     0.773     HIGH
                 0.780    0.375     0.675     0.780     0.724     0.409     0.773     0.742     LOW
Weighted Avg.    0.702    0.298     0.707     0.702     0.700     0.409     0.773     0.758
```

Fig. 10. Random Forest Classification Report

AUPRC score of 77.3% (see Fig.10). Unfortunately, Weka does not have a way to visualize the ensemble of trees model and so this model is difficult to interpret even though it performed the best.

The second best model was the one employing Logistic Regression to classify the samples. This model achieved an overall AUPRC score of 72.4% and a minority class AUPRC score of 72.2% (see Fig.11). Next, the model is interpreted by taking a look at the final coefficients associated with the different features to identify which attributes are driving the decision making of the model (see Fig.12). The insights from the coefficients are described below.

a) It is observed that many of the regions have a very high weighting (either negative or positive) which indicates that region plays a big role in the decision making of the model and this is in keeping with the first of our initial hypotheses.

b) The area of a country seems to have a big role to play. Since it is a positive value, the interpretation is that countries with a large land area are more vulnerable to climate related disasters. This could possibly be a case of simple probability, in that larger the land area of a country, greater the chance of at least some part of the country suffering from climate change.

c) Infant mortality rate has a somewhat high positive weighting, which suggests a slightly positive correlation between this feature and high-risk status. This is in keeping with the second of our initial two hypotheses.

d) GDP has a large negative coefficient. This means that as GDP decreases, the chances of a country being high-risk increases. Once again, this is aligned with our initial hypothesis that less developed and economically weaker countries are at greater risk to climate change.

e) The 'Crops' attribute seems to be contributing well to the decision making process as well. Unfortunately, a clear definition for this particular feature could not be obtained from the data source page and is therefore difficult to interpret.

```
Correctly Classified Instances         117.6284               67.2162 %
Incorrectly Classified Instances        57.3716               32.7838 %
Kappa statistic                          0.3443
Mean absolute error                      0.3571
Root mean squared error                  0.4835
Relative absolute error                 71.4242 %
Root relative squared error             96.7005 %
Total Number of Instances              175

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.604    0.260     0.699      0.604     0.648     0.348    0.733     0.722     HIGH
                 0.740    0.396     0.652      0.740     0.693     0.348    0.733     0.725     LOW
Weighted Avg.    0.672    0.328     0.675      0.672     0.671     0.348    0.733     0.724
```

Fig. 11. Logistic Regression Classification Report

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                                                          Class
Variable                                                   HIGH
============================================================
Region=ASIA (EX. NEAR EAST)                               3.537
Region=EASTERN EUROPE                                    -0.0327
Region=NORTHERN AFRICA                                  -18.8718
Region=SUB-SAHARAN AFRICA                                 1.1658
Region=LATIN AMER. & CARIB                                1.945
Region=C.W. OF IND. STATES                              -11.4318
Region=OCEANIA                                           -3.2733
Region=WESTERN EUROPE                                     3.2124
Region=NEAR EAST                                         -0.0711
Region=NORTHERN AMERICA                                 -29.1661
Region=BALTICS                                          -16.1898
Area (sq. mi.)                                           23.0634
Pop. Density (per sq. mi.)                               -3.3067
Coastline (coast/area ratio)                            -24.7593
Net migration                                            10.8395
Infant mortality (per 1000 births)                        1.9822
GDP ($ per capita)                                       -6.1829
Literacy (%)                                             -0.0671
Phones (per 1000)                                        -1.0364
Crops (%)                                                 6.3168
Climate                                                   0.7637
Birthrate                                                -1.5247
Deathrate                                                 0.5257
Agriculture                                              -1.9355
Intercept                                                -7.6853
```

Fig. 12. Logistic Regression Coefficient Values

8 CONCLUSION

The initial data visualization seemed to suggest that there was cause to believe the data would support the two hypotheses made in section 3. The clustering process then highlighted that the instances could indeed be grouped into two clusters, however these clusters did not perfectly distinguish between the low-risk and high-risk instances. The analysis of the clusters and their relationship with various features did show that there was good separation between the two clusters on the basis of infant mortality rate, GDP and region. This was also encouraging and suggested there may be truth to the hypotheses.

Finally, a supervised classification approach was used to identify the features that played an important role in determining whether a country was high-risk or low-risk. The coefficient values suggest a heavy dependence on the region attribute which strongly supports the first hypothesis. GDP had a strong negative correlation with the high-risk class which again strongly supports the second hypothesis (infant mortality rate supports the hypothesis as well although to a lesser extent).

Interestingly, the logistic regression model has highlighted a couple of other features that are influencing the decision making and it might be interesting to look into the logic behind this mathematical correlation to figure out what it means in the real world.

In conclusion, the data seems to support the two hypotheses made at the beginning of the study and the machine learning models verify the same to an extent. The analysis has shown that the geographical location of the country plays the most important role in determining its vulnerability to climate change and disasters associated with it. At the same time, the socioeconomic factors surrounding a country also play a role in how well they can deal with climate related disasters.

9 LIMITATIONS AND FUTURE WORK

The authors of the Climate Risk Index (CRI) analysis have mentioned that their work only takes into account the short term extreme weather events that affect a country. Not all the effects of climate change are this drastic or tangible. For instance, a phenomenon like global warming has been taking place for years and the increase in temperature is gradual but consistent. If the objective is to accurately model the adverse effects of climate change, the long term gradual effects will also have to be included in the study.

Additionally, the features of countries used in this study were quite basic. It might be worth including more complicated features such as longitude and latitude (more accurate geographical location based modelling) or including the Human Development Index (HDI) to get a direct representation of how developed a nation is.

References

[1] Aengenheyster, M., Feng, Q. Y., van der Ploeg, F., and Dijkstra, H. A.: The point of no return for climate action: effects of climate uncertainty and risk tolerance, Earth Syst. Dynam., 9, 1085–1095, https://doi.org/10.5194/esd-9-1085-2018, 2018.

[2] https://www.developmentaid.org/api/frontend/cms/file/2021/03/Global-Climate-Risk-Index-2021_1.pdf

[3] https://www.kaggle.com/fernandol/countries-of-the-world

[4] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

Appendix A: List of All Features in Dataset

1. Region
2. Population
3. Area (sq. mi.)
4. Pop. Density (per sq. mi.)
5. Coastline (coast/area ratio)
6. Net migration
7. Infant mortality (per 1000 births)
8. GDP ($ per capita)
9. Literacy (%)
10. Phones (per 1000)
11. Arable (%)
12. Crops (%)
13. Other (%)
14. Climate
15. Birthrate
16. Deathrate
17. Agriculture
18. Industry
19. Service
20. Risk (class labels)