REFRAIN*: PATTERNS AND PREDICTIVE POWER OF COUNTRY SOCIO-ADAPTIVE STATUS IN AN INTERNATIONAL MUSICAL SETTING

ISYE 6740: COMPUTATIONAL DATA ANALYTICS
SUMMER 2023
PROFESSOR XIE

Emily Bryans Dobar
Online Masters of Analytics
Georgia Institute of Technology
Atlanta, GA 30332
edobar3@gatech.edu

Contents

**Refrain* is the title of the first-ever winning song in Eurovision 1956 [1] and also references a repeating pattern or chorus in a song.

1 Background

Le Grand-Prix Eurovision de la Chanson Européenne, better-known as the Eurovision Song Contest, or more commonly as simply Eurovision, is an annual singing competition held by the European Broadcasting Union that includes entries from many countries across Europe, Western Asia, and Australia [12]. Since its founding in 1956, Eurovision has been broadcast as "celebration of European unity and culture" and, though it was founded in a post-war society with an intention of maintaining an apolitical standpoint, has always had its social and political undertones presented during performances, within the lyrics of the songs, or even during the hosting of the event [11].

While most political undertones tend towards war, such as countries in the Soviet Bloc not being allowed to participate during the Cold War or, most recently, Russia being banned due to its siege of Ukraine in 2022 [11], other factors such as cultural or social issues can come into play, with such examples as Ireland's 2018 entry supporting same-sex love and LGBTQ+ rights [8] and Serbia's 2022 entry commenting on healthcare in Serbia [10].

Another issue that is no stranger to the Eurovision stage is climate change, with Finland's notable 2018 entry speaking out in favor of fighting the global crisis [7]. Every year, climate change is pushing itself more and more onto society's radars; its impacts negatively affect almost every aspect of society from water scarcity and food production to human health and well-being to ecosystems, infrastructures, and economies [6]. And as time moves forward, the impacts of climate change become more unavoidable and devastating, especially to countries that are considered more vulnerable to the effects and are less prepared to combat them [13].

Different factors can be used to determine how vulnerable or prepared a country is to face climate change. The Notre Dame Global Adaptation Index (ND-GAIN) Country Index is an index that assesses both a country's vulnerability to the negative impacts of climate change, as well as its readiness to face climate change using indicators to calculate an overall adaptiveness score, including items for both vulnerability (Adaptive Capacity, Sensitivity, and Exposure) and readiness (Social, Economic, and Governance) [3].

In looking at the impacts of local and global socio-political issues on countries' entries into Eurovision each year, it may be possible to use the ND-GAIN Country Index to explore potential resulting patterns and groupings of countries and their performances in Eurovision over time, as well as using the index to build a predictive model to predict Eurovision outcomes based on a country's perceived adaptive power towards climate change. As Eurovision is an international stage, reaching not only viewers in participating countries but viewers globally, the socio-political statements and movements made during performances can be incredibly impactful outside the realm of music, and should climate change vulnerability and readiness have a significant impact on Eurovision performances and results, this could prompt more to be done on a global scale to combat climate change.

2 Research Questions

I intend to explore the relationships between climate change and Eurovision through cluster and regression analyses between the Notre Dame Global Adaptation Index Country Index and Eurovision results from 1999-2021, and since Eurovision 2020 was canceled, I also intend to explore the possible outcomes of the first-ever canceled Eurovision by using the final predictive model selected during analysis.

- Does a country's perceived adaptiveness towards climate change affect performance in the Eurovision Song Contest?
- Can the ND-GAIN Country Index be used to effectively predict Eurovision outcomes?
- What would Eurovision 2020 have looked like using this model?

3 Hypotheses

Initially, I hypothesize that a country's adaptiveness towards climate change does affect its performance in the Eurovision Song Contest. Further, I hypothesize that the ND-GAIN Index can effectively predict Eurovision outcomes and therefore give a reasonable prediction of the outcomes of Euorvision 2020, as compared to the predictions made regarding the odds of winning for each participating country in 2020 [5].

4 Data

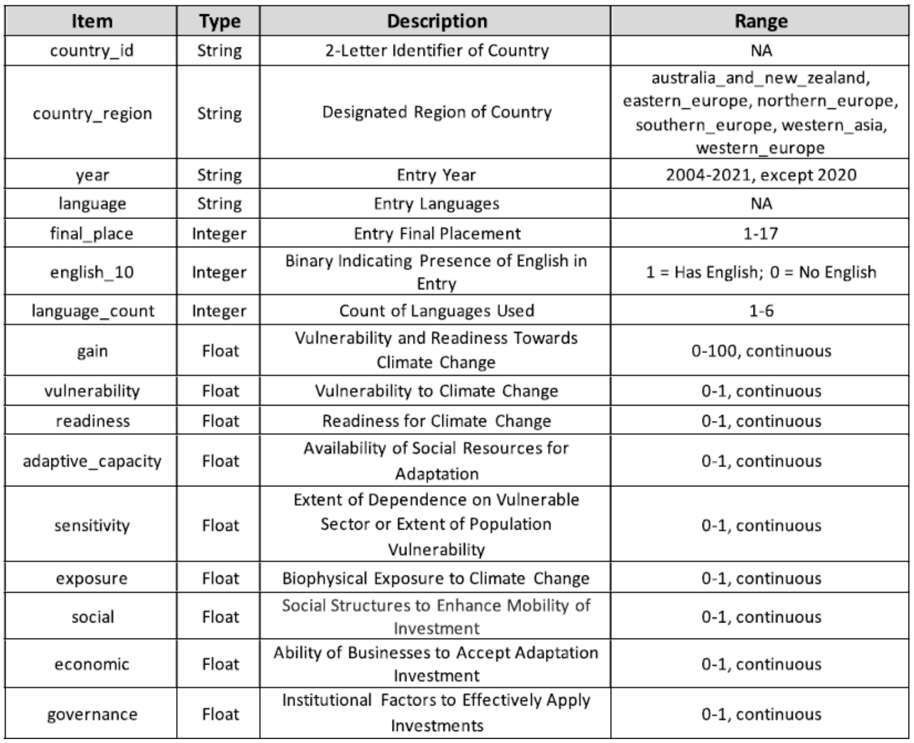
For this analysis, I used data from two separate datasets: a Eurovision dataset with all results from every Eurovision Song Contest from 1956-2021 [9] and the ND-GAIN Country Index datasets with scores for each indicator from 1995-2021 [3].

I merged the two datasets by country and year and kept only the data from 1999-2021 due to ND-GAIN only having data as far back as 1995 and since Eurovision changed submission rules in 1999 to allow submissions in any language rather than requiring submissions be in an official language of the country they are representing [2]. Reducing data to 1999 allowed me to include in the analysis a binary variable to indicate whether English is present in each submission, as well as allowed for some reduction of possible bias due to contest requirements. Similarly, I reduced the dataset moreso to 2004, as this would ensure all contests present in the data would include a semi-final, which allows for further reduction of bias due to contest structure.

For the analysis, I separated data from the canceled 2020 Eurovision, which was canceled due to the COVID-19 pandemic and has no official results/data [14], to use for additional analysis later. I also recoded the region variable to be values 1-6, with the numeric counterparts being representative of each region in alphabetic order, as well as removed all data points where the country did not qualify for the final (where final place = 0).

It is important to note that small countries, such as Andorra, Monaco, and San Marino, were removed due to lack of data for them. Romania was removed from the third analysis, as there is a lack of Governance data for it.

For a summary of the merged dataset, see Figure 1.



Item	Type	Description	Range
country_id	String	2-Letter Identifier of Country	NA
country_region	String	Designated Region of Country	australia_and_new_zealand, eastern_europe, northern_europe, southern_europe, western_asia, western_europe
year	String	Entry Year	2004-2021, except 2020
language	String	Entry Languages	NA
final_place	Integer	Entry Final Placement	1-17
english_10	Integer	Binary Indicating Presence of English in Entry	1 = Has English; 0 = No English
language_count	Integer	Count of Languages Used	1-6
gain	Float	Vulnerability and Readiness Towards Climate Change	0-100, continuous
vulnerability	Float	Vulnerability to Climate Change	0-1, continuous
readiness	Float	Readiness for Climate Change	0-1, continuous
adaptive_capacity	Float	Availability of Social Resources for Adaptation	0-1, continuous
sensitivity	Float	Extent of Dependence on Vulnerable Sector or Extent of Population Vulnerability	0-1, continuous
exposure	Float	Biophysical Exposure to Climate Change	0-1, continuous
social	Float	Social Structures to Enhance Mobility of Investment	0-1, continuous
economic	Float	Ability of Businesses to Accept Adaptation Investment	0-1, continuous
governance	Float	Institutional Factors to Effectively Apply Investments	0-1, continuous

Figure 1: Data Summary

5 Methods

I divided the dataset for three separate analyses using a country's final place in Eurovision as the dependent variable and the country's region, language count in the song, and whether or not English is one of the languages as independent variables alongside the following unique items from the ND-GAIN Country Index dataset for each of the three analyses:

i ND-GAIN Score

ii Vulnerability Score, Readiness Score

iii Adaptive Capacity, Sensitivity, Exposure, Social, Economic, Governance

The division of analysis stems from how the ND-GAIN Country Index is divided into separate indicators, and subsequently components and sectors, to calculate a country's overall ND-GAIN score. See Figure 2. To ensure as equal division of both vulnerability and readiness, the third analysis will focus on the main three components of each, rather than exploring the six sub-sectors of vulnerability.

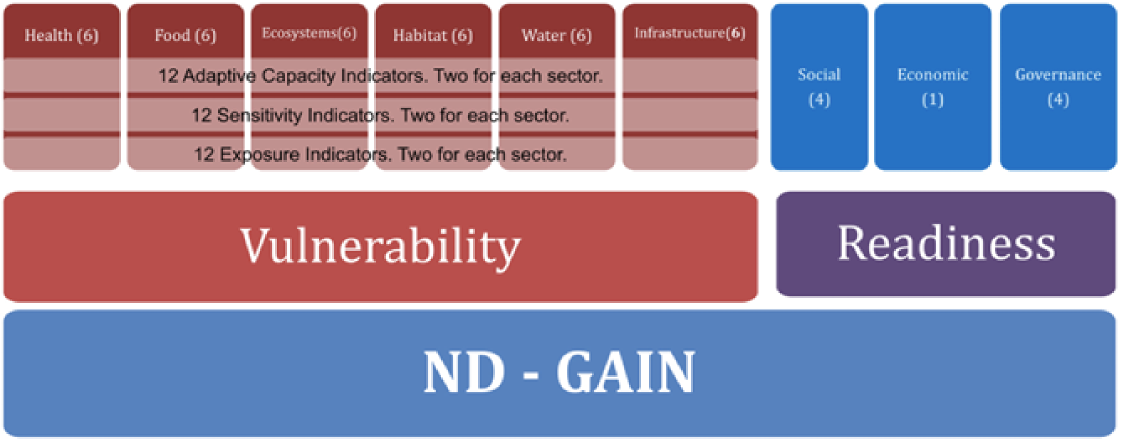

Health (6)
Food (6)
Ecosystems(6)
Habitat (6)
Water (6)
Infrastructure(6)
12 Adaptive Capacity Indicators. Two for each sector.
12 Sensitivity Indicators. Two for each sector.
12 Exposure Indicators. Two for each sector.
Social (4)
Economic (1)
Governance (4)
Vulnerability
Readiness
ND - GAIN

Figure 2: Summary of ND-GAIN Indicators[3]

For each of the analyses, I intend to perform cluster analysis and then build regression models to examine predictive power of each of the datasets' features in predicting performance in Eurovision. Each dataset was divided into training and test sets to train and test the models. To compare models, I calculated and compared R^2, root mean squared error (RMSE), and accuracy, which will be calculated by comparing the predictions on the test set from the model to the true values in the test set. The accuracy will also use a margin of 3, rather than focusing on exact matches, as the resulting placement in Eurovision isn't continuous.

5.1 Exploratory Data Analysis

Before the three analyses, I performed preliminary exploratory data analysis on the variables present in the dataset. Figure 3 shows the boxplots for each of the features in the data to examine the spread and potential outliers of each. Only *economic* and *sensitivity* seem to have some potential outliers outside of their respective boxplots. Otherwise, most features seem to be similarly-spread, save for *english_10* and *language_count* which are binary and categorical variables, respectively.

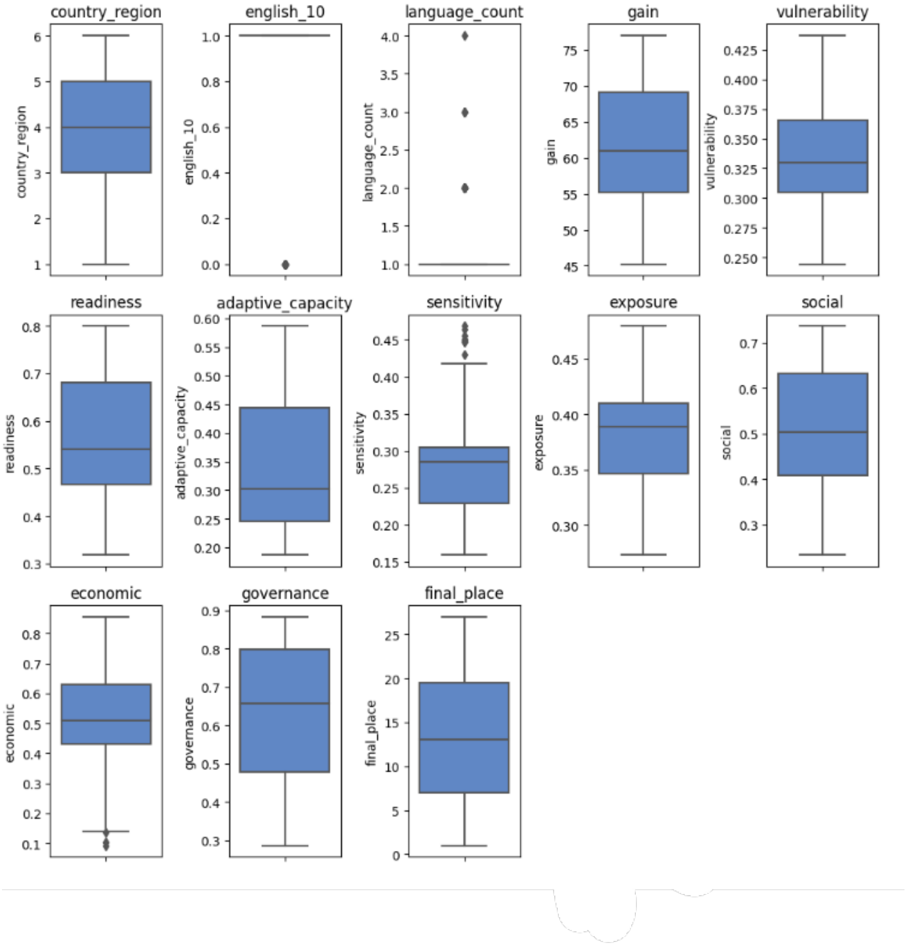


Figure 3: Boxplot of Features

Figure 4 shows the histograms of each of the features in the data to observe the distribution of each. Many of the numerical features appear to be normally distributed, with only *adaptive_capacity*, *social*, and *governance* demonstrating some skewing. The response variable, *final_place* does not look very normally distributed, but this is more likely due to how the placements in Eurovision may fall, as they would not appear in a bell curve naturally.

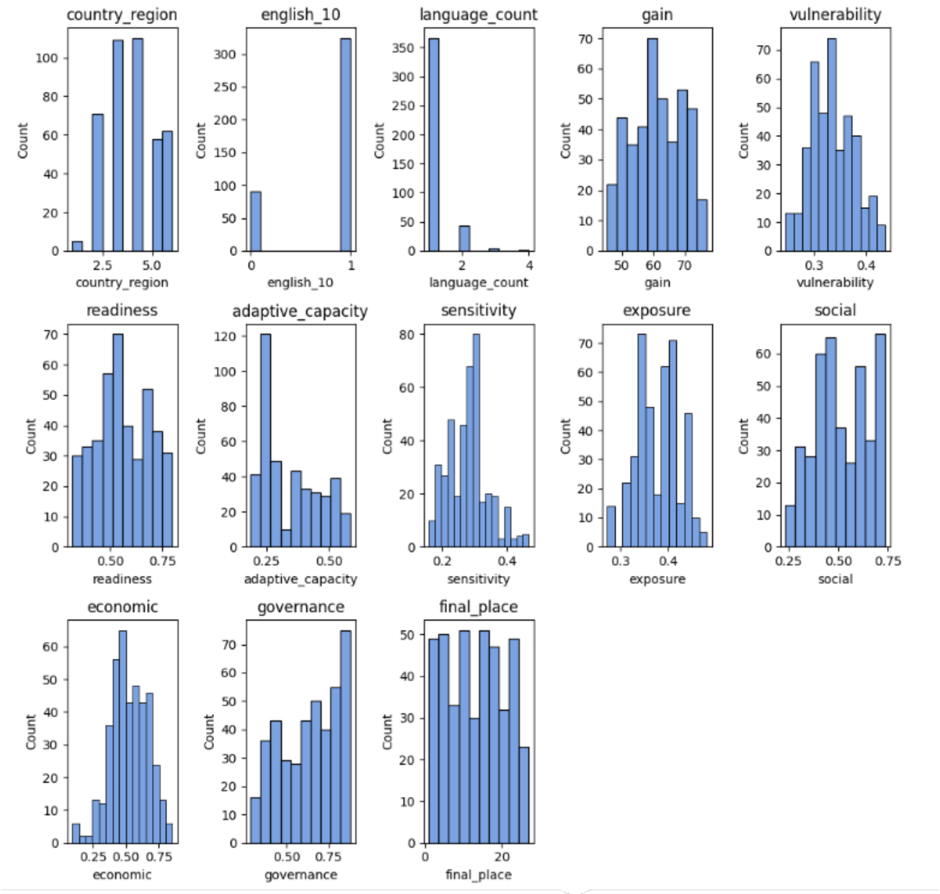


Figure 4: Histogram of Variables

To examine the bivariate relationships between the features in each of the three datasets, I looked at scatterplots between each of the variables.

For the ND-GAIN Score dataset, Figure 5 shows the bivariate relationships between the features. There do not appear to be any obvious linear relationships between any of the variables. However, *region* and *gain* seem to have some mild relationship, as certain regions tend to have higher GAIN scores than others, which makes sense given how GAIN is calculated, since certain regions of the world may have more resources or vulnerability towards climate change. *final_place* and *gain* appear to have a random relationship, as there are no obvious trends in the scatterplot.

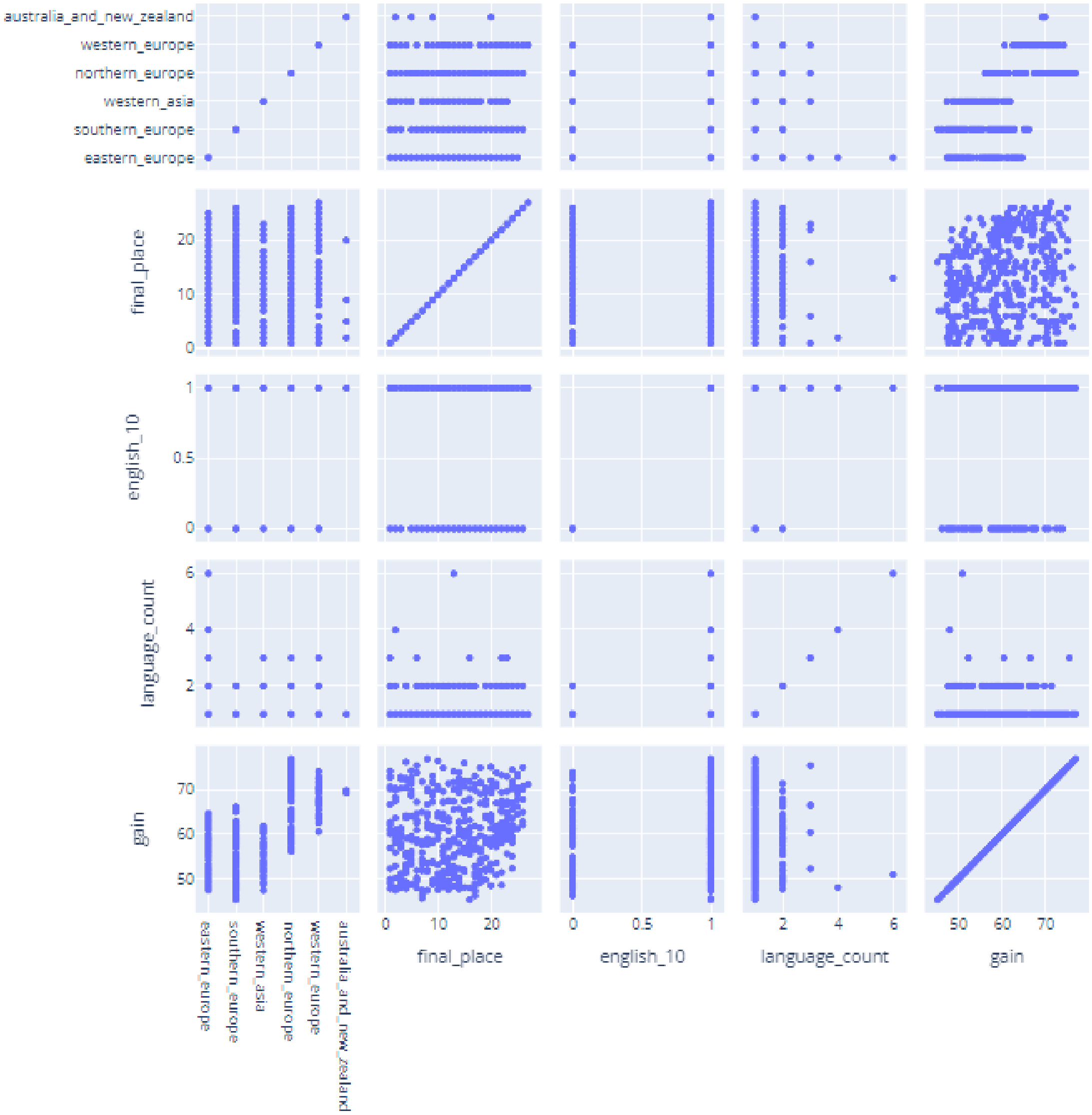


Figure 5: Bivariate Relationships Between Features (ND-GAIN Score)

For the ND-GAIN Indicators dataset, Figure 6 shows the bivariate relationships between the features. Similarly, as expected, *region* seems to have a relationship with both *vulnerability* and *readiness*, and *final_place* still doesn't seem to have any discernible relationship with any of the predictors. *readiness* and *vulnerability* seem to have a negative linear relationship, which makes sense given the description of each of those features.

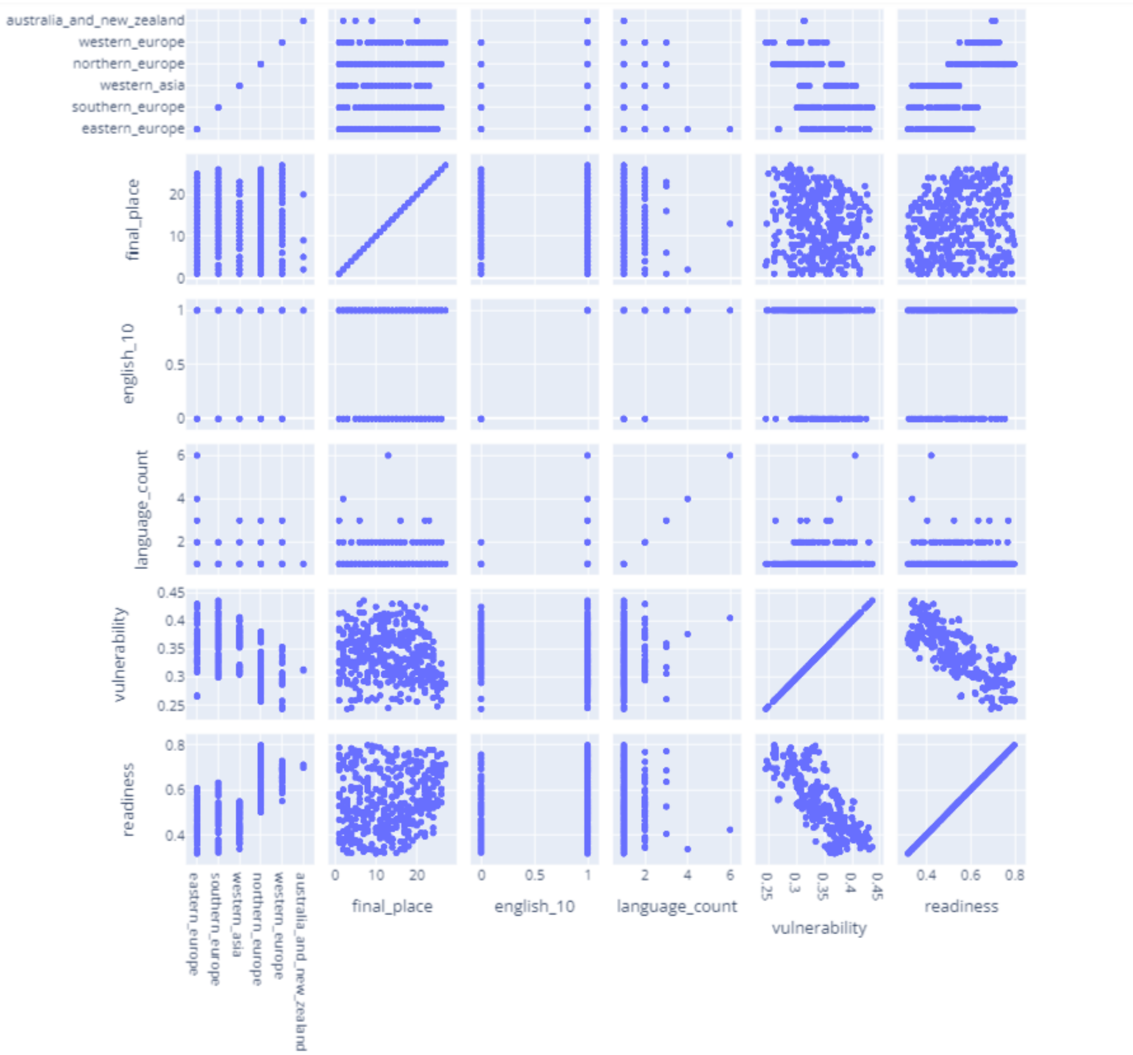


Figure 6: Bivariate Relationships Between Features (ND-GAIN Indicators)

For the ND-GAIN Indicator Components dataset, Figure 7 shows the bivariate relationships between the features. As in the two previous figures, *final_place* does not seem to have any discernible relationship with any of the predictor variables. Some of the ND-GAIN Indicator Components seem to have potential linear or non-linear relationships.

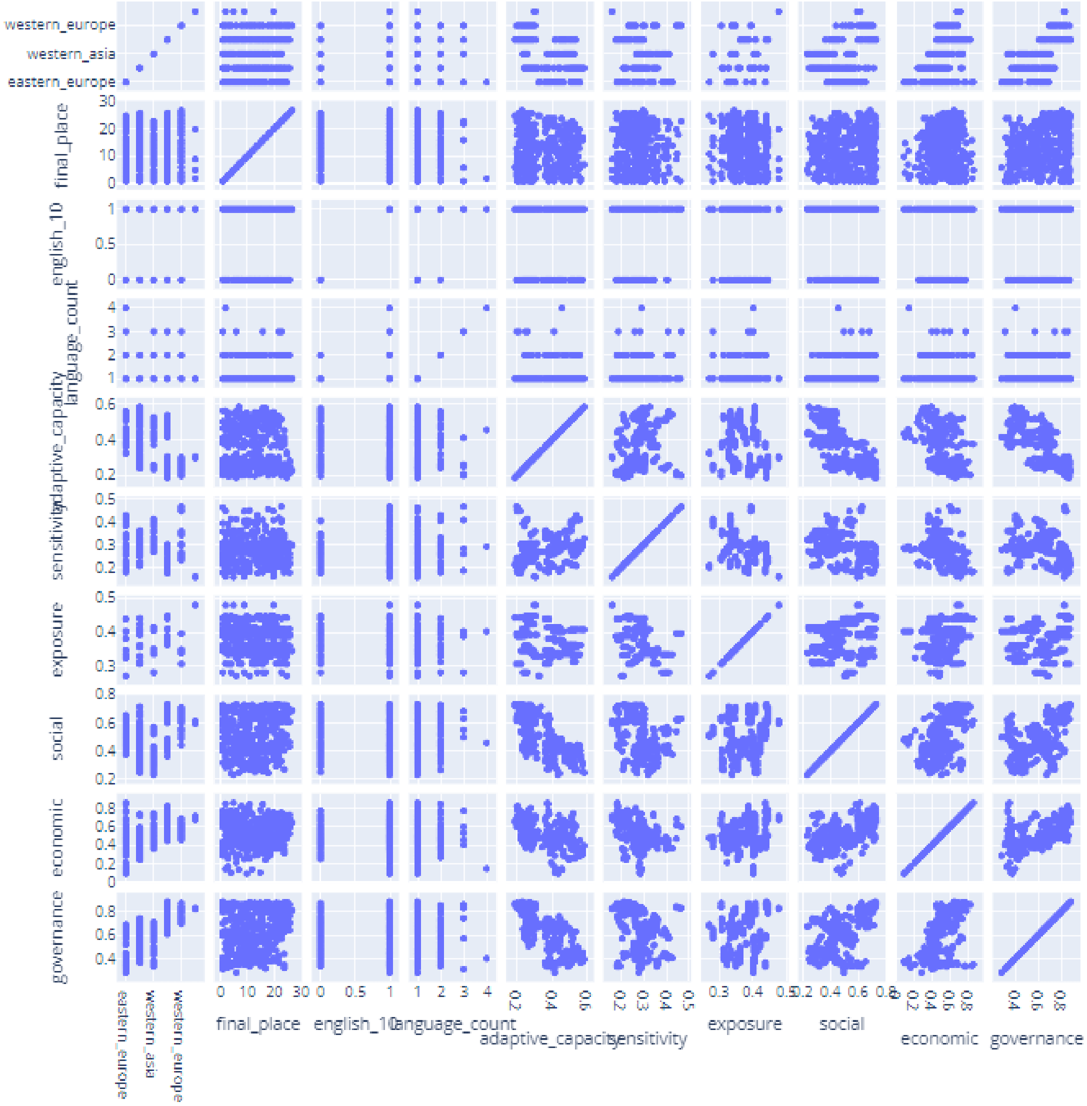


Figure 7: Bivariate Relationships Between Features (ND-GAIN Indicator Components)

Lastly, I examined the correlations between each of the features across the full dataset. In looking at Figure 8, ignoring the ND-GAIN features that are expected to have stronger correlations such as *vulnerability* and *readiness*, most features do not have strong correlations. The features with the strongest correlations with the response variable *final_place* are *governance* (0.22), *vulnerability* (-0.18), and *gain* (0.17).

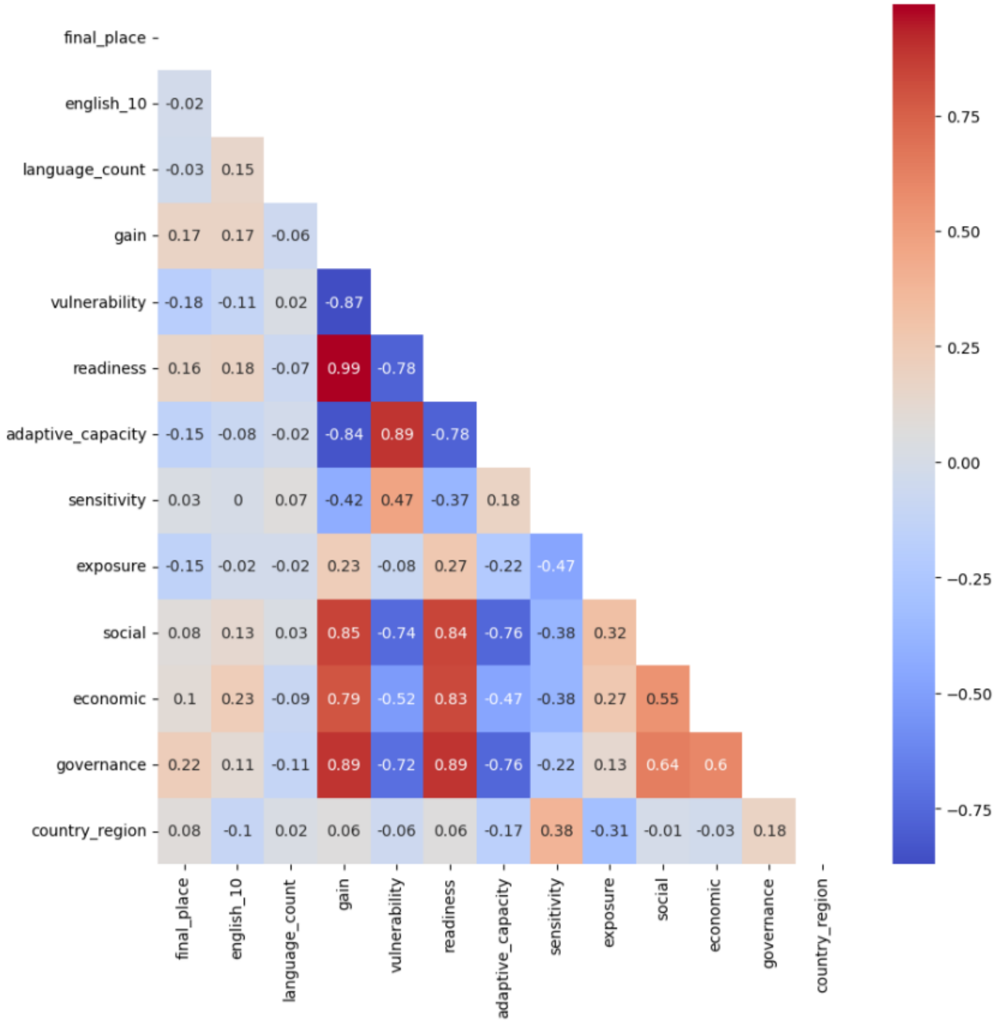


Figure 8: Correlations of Variables

5.2 ND-GAIN Score

The following analysis is exploring the relationship between the overall ND-GAIN score of each country and its final placement in Eurovision.

5.2.1 Cluster Analysis

Figure 9 visualizes *gain* versus *final_place*, with color being determined by *region* and size being determined by *language_count*. While there are no obvious relationships visible between GAIN score and final placement in Eurovision, the different regions somewhat form clusters along the x-axis.

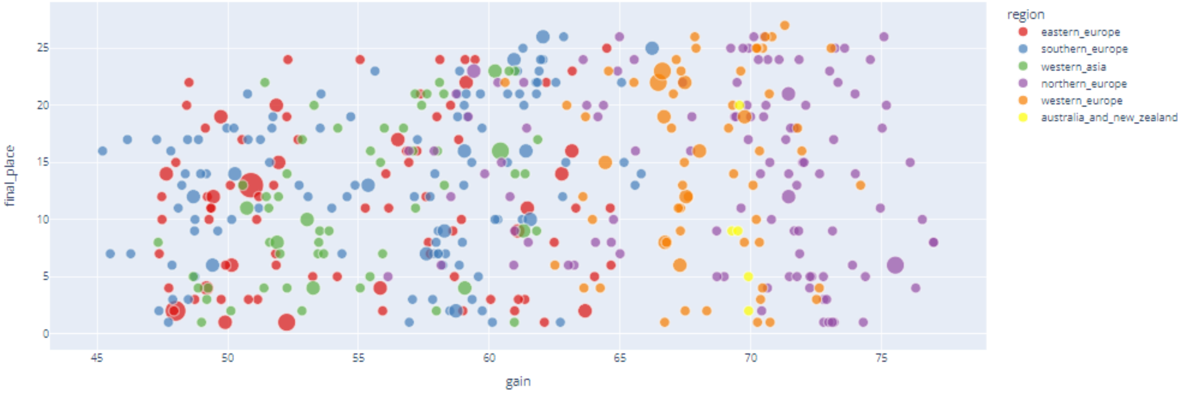


Figure 9: Scatterplot of Features

After using the elbow method to find that the optimal number of clusters is 3, I performed k-means clustering to reveal the 3 clusters present in the data. Figure 10 shows a polar plot of the three clusters and indicates the strength of each feature in determining the cluster placement of each data point. The strongest feature for the red and green clusters in the plot is *english_10*, which suggests that the inclusion of English in the song is a strong predictor in determining data points in those two clusters. The smallest cluster, the blue cluster, has *region* as its strongest predictor. Interestingly, *language_count* is not a strong predictor for any of the clusters, while *final_place* and *gain* both seem to have moderate influence in cluster assignment.

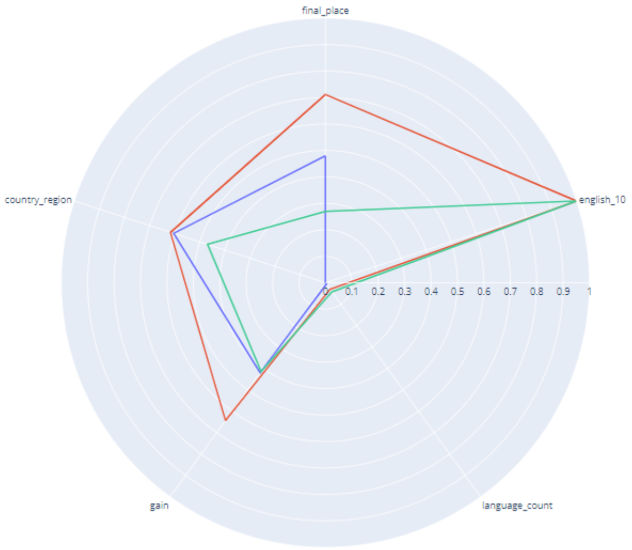


Figure 10: Polar Plot of Clusters

Figure 11 visualizes the cluster placements of each of the data points in a scatterplot with GAIN Score versus final Eurovision placement. The yellow cluster, the largest cluster, contains countries with lower GAIN scores regardless of Eurovision performance. The other two clusters both contain countries with higher GAIN scores, but the green cluster contain those that place lower in Eurovision, while the purple cluster contains those that place higher in Eurovision.

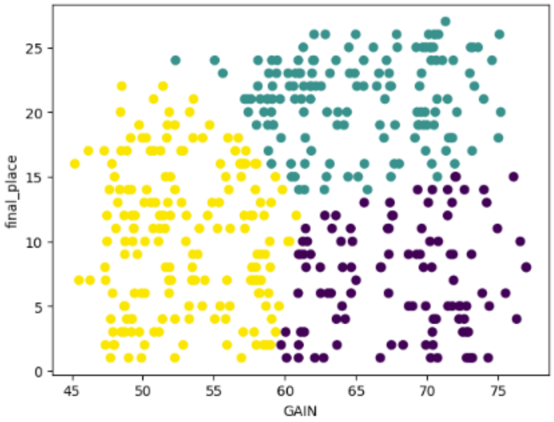


Figure 11: Scatterplot of Clusters

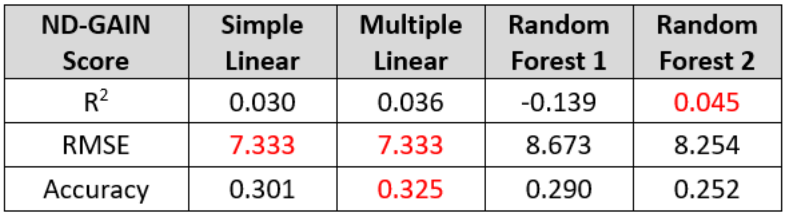
5.2.2 Regression Analysis

In using the overall ND-GAIN Score to predict performance in Eurovision, I built four different regression models to compare performance: a simple linear model, a multiple linear model, a random forest model with one predictor, and a random forest model with multiple predictors. The simple linear model and the first random forest model both only include the GAIN Score as predictors, while the multiple linear model and the second random forest model both include the GAIN Score, as well as the other features in the dataset indicating if English is included in the song, the number of languages included in the song, and the region of the country. Figure 12 shows the results of each of the regression models, with red indicating the best values for each of the model performance metrics.

Overall, none of the models performed very well on the data, with the highest R^2 value achieved being $= 0.045$, suggesting that a maximum of 4.5% of the variance in the data is explained by the second random forest regression model. However, the lowest RMSE was $= 7.333$, which was resulting from both the simple and multiple linear models. The most accurate model when using the test dataset was the multiple linear model, with an accuracy of 32.5%. The first random forest model did not perform best in terms of any metric and is therefore eliminated as a possible model for this dataset.

While the second random forest model had the highest R^2, the multiple linear model seems to be the best-performing model out of the four to predict Eurovision performance using the overall ND-GAIN Score. The multiple linear model had a similar, even if slightly lower, R^2 to the second random forest model, but it had a better RMSE and was more accurate in predictions.

For full model summaries and trimmed trees, see the Appendix Figures 28 through 31.



ND-GAIN Score	Simple Linear	Multiple Linear	Random Forest 1	Random Forest 2
R^2	0.030	0.036	-0.139	0.045
RMSE	7.333	7.333	8.673	8.254
Accuracy	0.301	0.325	0.290	0.252

Figure 12: Model Comparison (ND-GAIN Score)

5.3 ND-GAIN Indicators

The following analysis is exploring the relationship between the ND-GAIN indicator scores for climate change vulnerability and readiness of each country and its final placement in Eurovision.

5.3.1 Cluster Analysis

Figures 13 and 14 visualize *vulnerability* versus *final_place* and *readiness* versus *final_place*, respectively, with color being determined by *region* and size being determined by *language_count*. While there are no obvious relationships visible between readiness and final placement in Eurovision, there do seem to be some small clusters forming between vulnerability and final placement in Eurovision.

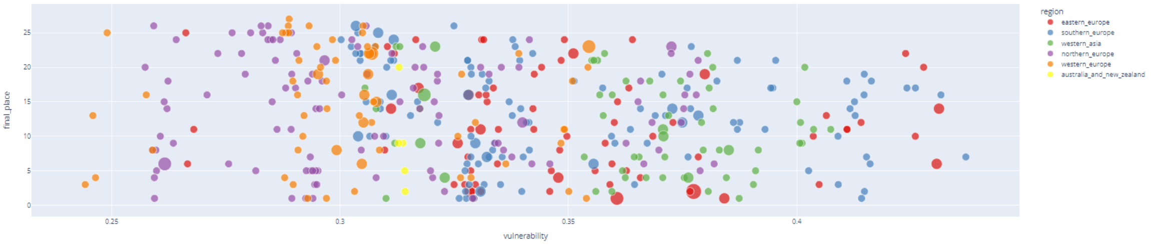


Figure 13: Scatterplot of Features (Vulnerability)

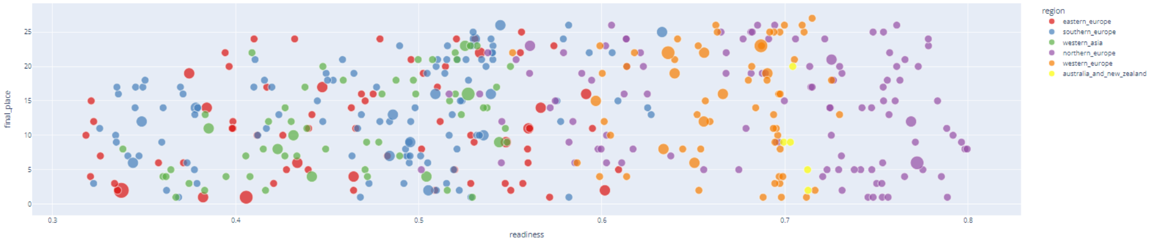


Figure 14: Scatterplot of Features (Readiness)

After using the elbow method to find that the optimal number of clusters is 3, I performed k-means clustering to reveal the 3 clusters present in the data. Figure 15 shows a polar plot of the three clusters and indicates the strength of each feature in determining the cluster placement of each data point. The strongest feature for the red and green clusters in the plot is *english_10*, which suggests that the inclusion of English in the song is a strong predictor in determining data points in those two clusters, which is similar to the polar plot in Figure 10. The smallest cluster, the blue cluster, has *region* as its strongest predictor, which is also mimicking behaviors in the previous analysis. Interestingly, *language_count* is still not a strong predictor for any of the clusters. However, the red cluster seems to be more determined by *readiness*, while the blue and green clusters are more determined by *vulnerability*.

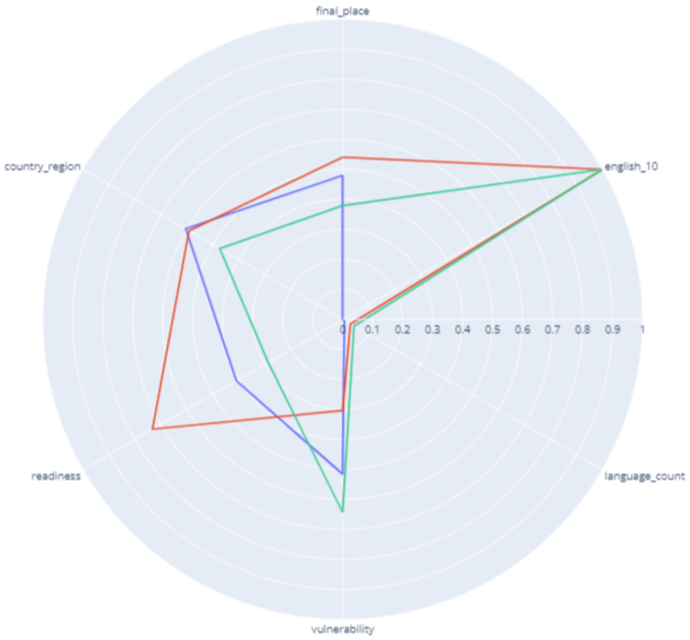


Figure 15: Polar Plot of Clusters

Figure 16 visualizes the cluster placements of each of the data points in a 3-D scatterplot with vulnerability and readiness towards climate change on the x and y axes, while Eurovision performance is on the z. The clusters all seem to have similar range in terms of Eurovision performance, which indicates that readiness and vulnerability towards climate change may be more indicative of clustering than the overall GAIN Score.

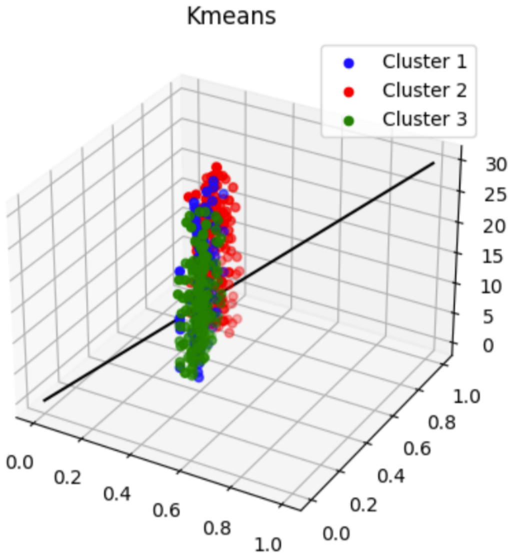


Figure 16: Plot of Clusters

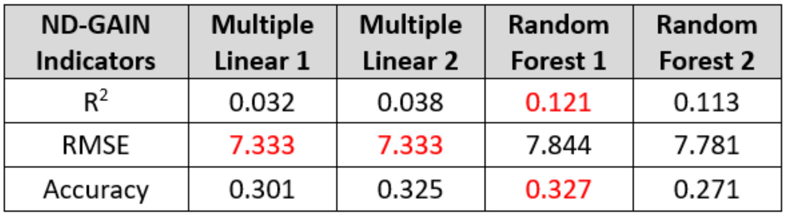
5.3.2 Regression Analysis

In using the ND-GAIN Indicators to predict performance in Eurovision, I built four different regression models to compare performance: two multiple linear models and two random forest models. The first of each type of model only include the ND-GAIN Indicators, while the second of each type of model include the ND-GAIN Indicators, as well as the other features in the dataset indicating if English is included in the song, the number of languages included in the song, and the region of the country. Figure 17 shows the results of each of the regression models, with red indicating the best values for each of the model performance metrics.

Overall, none of the models performed very well on the data, with the highest R^2 value achieved being $= 0.121$, suggesting that a maximum of 12.1% of the variance in the data is explained by the first random forest regression model. However, the lowest RMSE was $= 7.333$, which was resulting from both of the multiple linear models. The most accurate model when using the test dataset was the first random forest, with an accuracy of 32.7%. The second random forest model did not perform best in terms of any metric and is therefore eliminated as possible models for this dataset.

While both multiple linear models had the lowest RMSE values, the first random forest model seems to be the best-performing model out of the four to predict Eurovision performance using the ND-GAIN Indicators with the highest R^2 and accuracy values.

For full model summaries and trimmed trees, see the Appendix Figures 32 through 35.



ND-GAIN Indicators	Multiple Linear 1	Multiple Linear 2	Random Forest 1	Random Forest 2
R^2	0.032	0.038	0.121	0.113
RMSE	7.333	7.333	7.844	7.781
Accuracy	0.301	0.325	0.327	0.271

Figure 17: Model Comparison (ND-GAIN Indicators)

5.4 ND-GAIN Indicator Components

The following analysis is exploring the relationship between a country's final placement in Eurovision and the following ND-GAIN indicator components: adaptive capacity, sensitivity, exposure, social, economic, governance

5.4.1 Cluster Analysis

Figures 18 through 23 show each of the indicator components *final_place*, with color being determined by *region* and size being determined by *language_count*. Many of the indicator components seem to show more evidence of relationships and clusters in the scatterplots than in the two previous analyses.

Figure 18 below is showing some potential clustering in terms of *adaptive_capacity*, with one cluster of points below 0.35, another between 0.35 and 0.45, and another above 0.45.

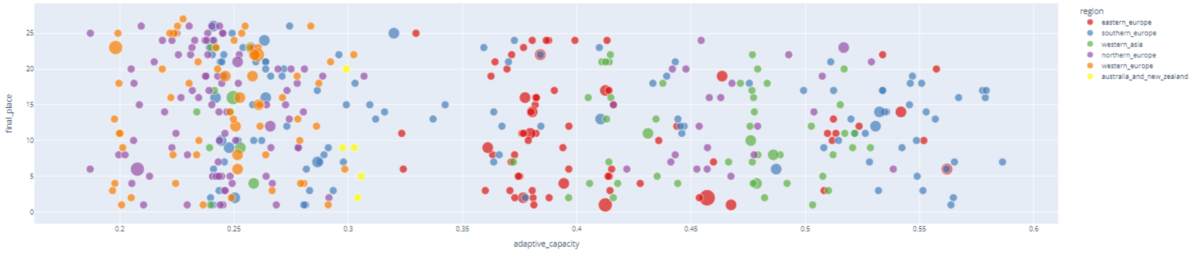


Figure 18: Scatterplot of Features (Adaptive Capacity)

Figure 19 is showing some clustering in terms of *sensitivity*, though not as well-defined as in some other predictors. One more-defined cluster is focused around 0.3, while other clusters may be between 0.15 and 0.25 and then above 0.35.

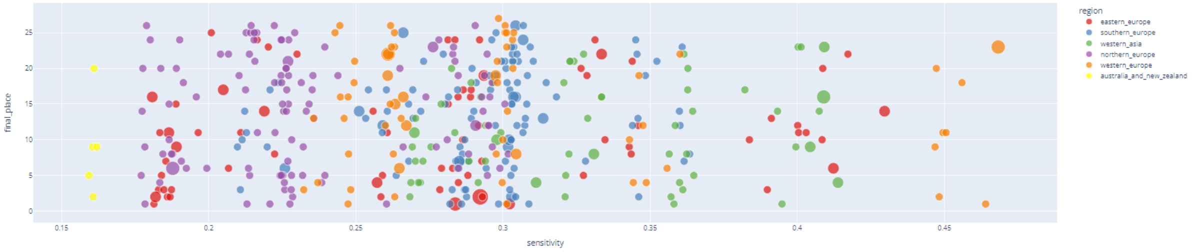


Figure 19: Scatterplot of Features (Sensitivity)

Figure 20 is showing about four different clusters appearing in terms of *exposure*: around 0.3, around 0.35, around 0.4, and around 0.45. Something else interesting to note is that the points in this plot seem more stacked, almost like a histogram, rather than randomly scattered like in other plots.

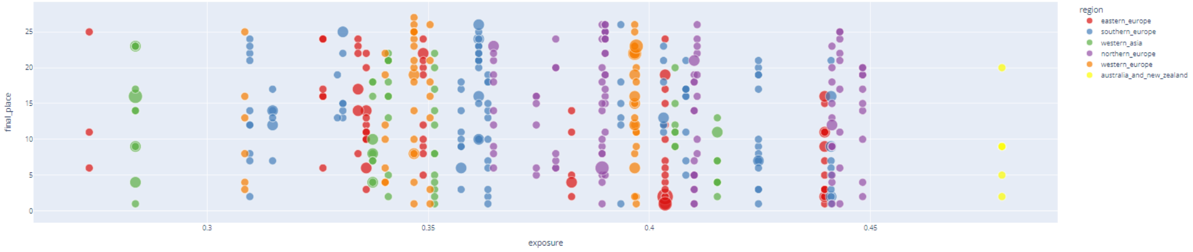


Figure 20: Scatterplot of Features (Exposure)

Figure 21 is showing more randomization, similar to the scatterplots in previous analyses. However, the regions are showing some clustering along the x-axis, which makes sense given certain regions and countries may align similarly in terms of social aspects.

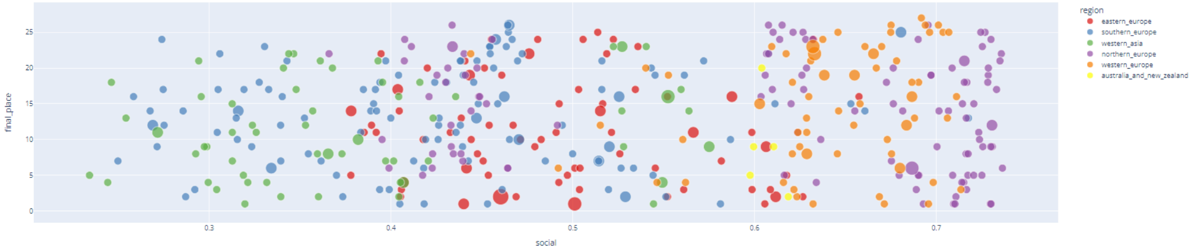


Figure 21: Scatterplot of Features (Social)

Figure 22 too is showing more randomization and not indicating any explicit clusters, though certain regions are still clustering together. However, there is some mild clustering between 0.4 and 0.5.

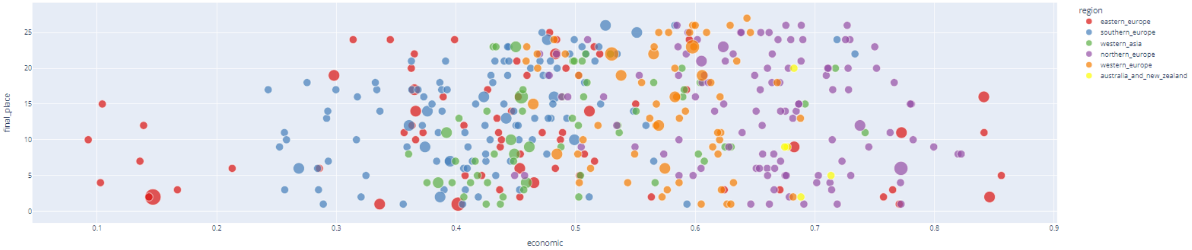


Figure 22: Scatterplot of Features (Economic)

Figure 23 is also not indicating any significant clusters, but this plot is showing more division and clustering in terms of region than others that show more overlap, which makes sense given that certain regions and countries may align more in terms of governance.

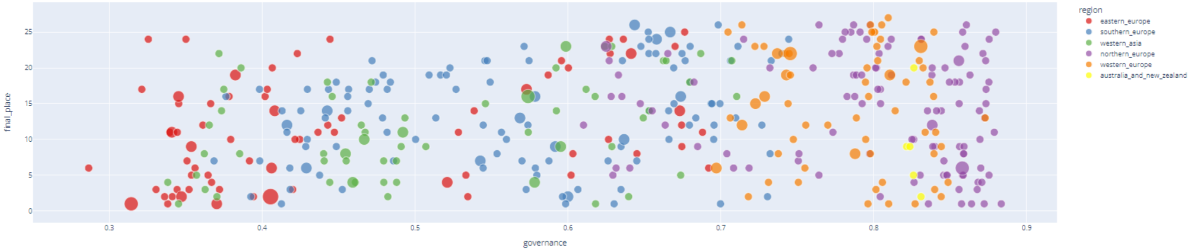


Figure 23: Scatterplot of Features (Governance)

After using the elbow method to find that the optimal number of clusters is 3, I performed k-means clustering to reveal the 3 clusters present in the data. Figure 24 shows a polar plot of the three clusters and indicates the strength of each feature in determining the cluster placement of each data point.

The strongest feature for the red and green clusters in the plot is *english_10*, which suggests that the inclusion of English in the song is a strong predictor in determining data points in those two clusters, which is similar to the polar plots for both previous analyses. The smallest cluster, the blue cluster, is more rounded than the other two, indicating similar strength across many different features. The red cluster generally has a stronger relationship with more features than he other two. The green and blue clusters both lean more towards *adaptive_capacity* than the red. Interestingly, *language_count* is still not a strong predictor for any of the clusters.

Due to the complexity of this dataset, visualizing the clustering on a 2-D or 3-D scatterplot was not reasonable or conducive to interpretation.

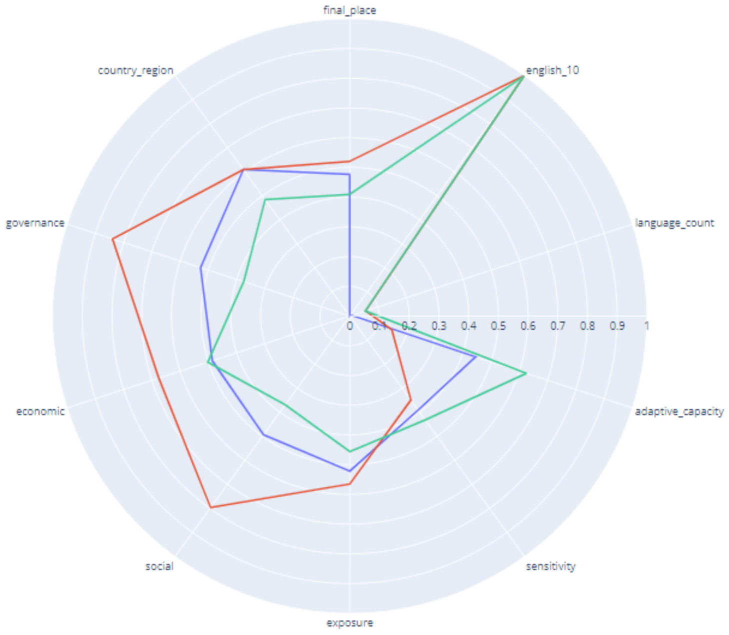


Figure 24: Polar Plot of Features

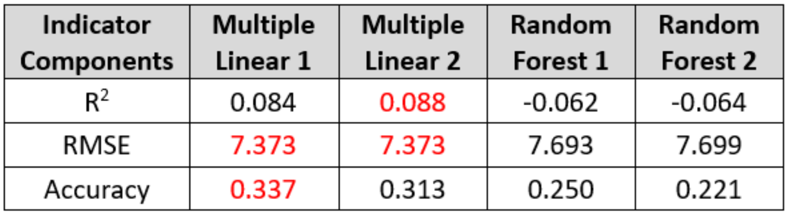
5.4.2 Regression Analysis

In using the ND-GAIN Indicator Components to predict performance in Eurovision, I built four different regression models to compare performance: two multiple linear models and two random forest models. The first of each type of model only include the ND-GAIN Indicator Components, while the second of each type of model include the ND-GAIN Indicator Components, as well as the other features in the dataset indicating if English is included in the song, the number of languages included in the song, and the region of the country. Figure 25 shows the results of each of the regression models, with red indicating the best values for each of the model performance metrics.

Overall, none of the models performed very well on the data, with the highest R^2 value achieved being $= 0.088$, suggesting that a maximum of 8.8% of the variance in the data is explained by the second multiple linear model. However, the lowest RMSE was $= 7.373$, which was resulting from both of the multiple linear models. The most accurate model when using the test dataset was the first multiple linear model, with an accuracy of 33.7%. Neither random forest model performed best in terms of any metric and are therefore eliminated as possible models for this dataset.

While the second multiple linear model had the highest R^2 and was tied for lowest RMSE, the first multiple linear model seems to be the best-performing model out of the four to predict Eurovision performance using the ND-GAIN Indicators. The multiple linear model had a similar, even if slightly lower, R^2 to the second multiple linear model and identical RMSE, but it was more accurate in predictions.

For full model summaries and trimmed trees, see the Appendix Figures 36 through 39.



Indicator Components	Multiple Linear 1	Multiple Linear 2	Random Forest 1	Random Forest 2
R^2	0.084	0.088	-0.062	-0.064
RMSE	7.373	7.373	7.693	7.699
Accuracy	0.337	0.313	0.250	0.221

Figure 25: Model Comparison (ND-GAIN Indicator Components)

6 Evaluation

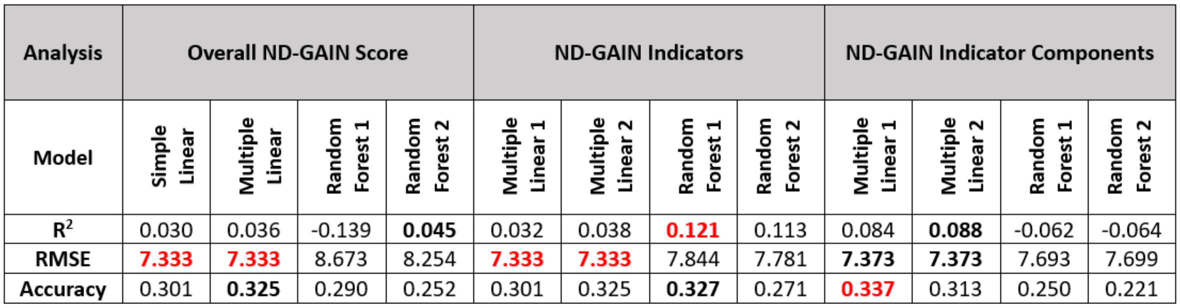
Comparing the clusters across each of the analyses, there were many similarities. The strongest featur in each analysis was *english_10*, while none of the analyses had *language_count* as a significant feature in determining clusters. With the most features, the ND-GAIN Indicator Components analysis seemed to have the most well-rounded clusters with each of the clusters having similar relationships to each of the features, just to different amplifications, such as red having strongest across most and blue having weakest across most. The ND-GAIN Indicators analysis seemed more polarized, with two of the clusters leaning towards *vulnerability* and one towards *readiness*. The ND-GAIN Score analysis had the most difference in sizes of clusters, as one cluster was significantly larger than the others, while another was much smaller and more-condensed.

Overall, it looks like clusters are more determined by the presence of English in the entry songs, as well as the country's region than determined by the ND-GAIN features. However, it appears that the ND-GAIN Indicators may have stronger cluster assignment power than the overall GAIN Score or the Indicator Components.

Figure 26 shows each of the analyses' regression model outputs together for a final comparison across each model in each analysis.

Across all 12 models, none performed very well on the data, with the highest R^2 value achieved being $= 0.121$, suggesting that a maximum of 12.1% of the variance in the data is explained by the first random forest model using the ND-GAIN Indicators. However, the lowest RMSE was $= 7.333$, which was resulting from four models: the simple linear model using the GAIN Score, the multiple linear model using the GAIN Score, and both of the multiple linear models using the ND-GAIN Indicators. The most accurate model when using the test dataset was the first multiple linear model using the ND-GAIN Indicator Components, with an accuracy of 33.7%. Most of the random forest models did not perform best in terms of any metric and are therefore eliminated as possible final models.

The ND-GAIN Indicators models had the most success in terms of overall metrics, even though the Indicator Components models had marginally better accuracy. Due to the model having the highest R^2 and second-best accuracy, the first random forest model using the ND-GAIN Indicators is selected as the best-performing model of the 12.



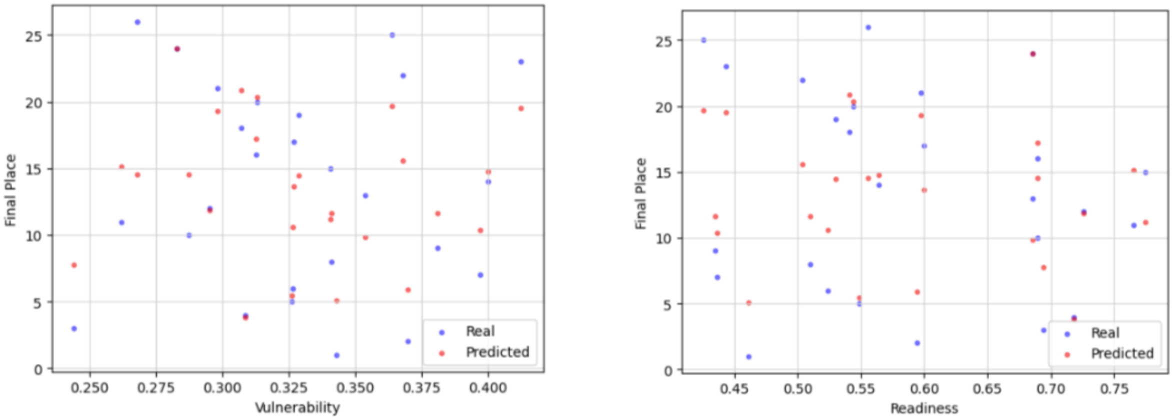
Analysis	**Overall ND-GAIN Score**				**ND-GAIN Indicators**				**ND-GAIN Indicator Components**			
Model	**Simple Linear**	**Multiple Linear**	**Random Forest 1**	**Random Forest 2**	**Multiple Linear 1**	**Multiple Linear 2**	**Random Forest 1**	**Random Forest 2**	**Multiple Linear 1**	**Multiple Linear 2**	**Random Forest 1**	**Random Forest 2**
R^2	0.030	0.036	-0.139	**0.045**	0.032	0.038	**0.121**	0.113	0.084	**0.088**	-0.062	-0.064
RMSE	**7.333**	**7.333**	8.673	8.254	**7.333**	**7.333**	7.844	7.781	**7.373**	**7.373**	7.693	7.699
Accuracy	0.301	**0.325**	0.290	0.252	0.301	0.325	**0.327**	0.271	**0.337**	0.313	0.250	0.221

Figure 26: Model Comparison (All Models)

6.1 Eurovision 2020: What If?

After selecting the best-performing model out of the 12, the first random forest model using the ND-GAIN Indicators, I used it to compare predictions for the 2020 Eurovision Song Contest, which was canceled due to the COVID-19 Pandemic [14].

Since this contest does not have true results, I compared the model predictions to official predictions as to the results of the 2020 Eurovision Song Contest, based on bookmaker odds predicting the chances of winning for each country in that year's contest [5]. Figure 27 shows the prediction comparisons using 2-D scatterplots with both vulnerability and readiness versus final performance in Eurovision. As seen in the figures, the real (which are the predictions made by the bookmaker odds) and the predicted (which are output from the model) values tend to fall closely together, suggesting somewhat decent predictions, though not perfect.



(a) Vulnerability

(b) Readiness

Figure 27: Model Performance on 2020 Data

After running the model and looking at the comparative figures, I also calculated the accuracy of the model in predicting the 2020 final performance based on the bookmaker odds, with a margin of 3 since the model is attempting to predict non-continuous values. Many of the predicted placements were very close to the real placements, with a total accuracy of approximately 39%, which, while not perfect, is on par with the model's accuracy during testing.

While the models did not agree perfectly on which country would win, they did agree that the top five for 2020 would contain Bulgaria, Lithuania, Switzerland, Iceland, and Russia.

For a full list of real versus predicted final placement values, see the Appendix Figure 40.

7 Future Work

In the future, expanding upon this exploration, I would like to explore different types of models. This data typically performed better with more variables and with the non-parametric models, which suggests that the more-linear models may not be best-fit for this kind of data. What also suggests this is that the final placement in Eurovision isn't a continuous variable, which makes models that predict continuous variables not perform as well. With that information, I tended to rely more on accuracy than R^2 in examining model performance, so with less-linear models, the R^2 could improve significantly.

Doing an exploration of the data in terms of logistic regression to predict probability of qualifying in the finale or in placing in a certain percentile in the finale could be more efficient as well, as this would reduce the error and bias added by the response variable being continuous in the linear models and likely improve R^2 as well.

Something else to explore in the future would be feature selection. Many features seemed to be more or less influential during cluster analysis, such as *english_10* and *language_count*, and alongside different variables appearing to be statistically significant in the model outputs and different variables selected during the random forest regression, this suggests that the models could further be improved simply by way of a feature selection process and rebuilding the models as-is with the new set of features.

Lastly, it could be worth looking into the sectors that fall within vulnerability in the ND-GAIN Country Index. I focused on indicators and components in this analysis for equal consideration and division of the data as I broke the analysis down into three stages, but seeing as there are further ways of dividing the data, this could prove both interesting and possibly informative with more information and exploration in the future.

References

[1] Assia, L., Voimard, G., Gardaz, É. (1956). Refrain [Recorded by performer L. Assia]. On D 18 265 [Vinyl, 7"]. Lugano, CH: Decca. (1956).

[2] Carlson, C. (2022, January 16). *An exploration of language*. Eurovisionworld. https://eurovisionworld.com/esc/an-exploration-of-language

[3] Chen, C., Noble, I., Hellman, J., Murillo, M., & Chawla, N. (2015, November). *Country index // Notre Dame Global Adaptation Initiative // University of Notre Dame*. Notre Dame Global Adaptation Initiative. https://gain.nd.edu/our-work/country-index/

[4] European Broadcasting Union (EBU). (2022, July 12). *The Origins of Eurovision*. Eurovision Song Contest. https://eurovision.tv/history/origins-of-eurovision

[5] *Odds eurovision song contest 2020*. Eurovisionworld. (2020, March 18). https://eurovisionworld.com/odds/eurovision-2020

[6] Intergovernmental Panel on Climate Change (IPCC). (2023). *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 3056 pp., https://doi.org/10.1017/9781009325844

[7] Keating, D. (2023, May 10). *The climate change messages behind the Eurovision Song Contest*. Energy Monitor. https://www.energymonitor.ai/features/the-climate-change-messages-behind-the-eurovision-song-contest/

[8] Kelleher, P. (2022, August 12). *17 of the biggest, brightest and queerest moments in Eurovision's joyful history*. PinkNews. https://www.thepinknews.com/2022/08/12/eurovision-lgbtq-conochita-wurst-dana-international-loreen/

[9] Levites, M., Wolfe, N., Willemse, J., Dudepoints, M, D., & Zammarelli, C. (2022, February 13). *Data*. ESC in Context. https://escincontext.com/resources/data/

[10] Macdonald, K. (2022, May 16). *Serbia's viral Eurovision Song featured Allegri's Miserere, and you might have missed it*. Classic FM. https://www.classicfm.com/composers/allegri/in-corpore-sano-konstrakta-eurovision-miserere-serbia/

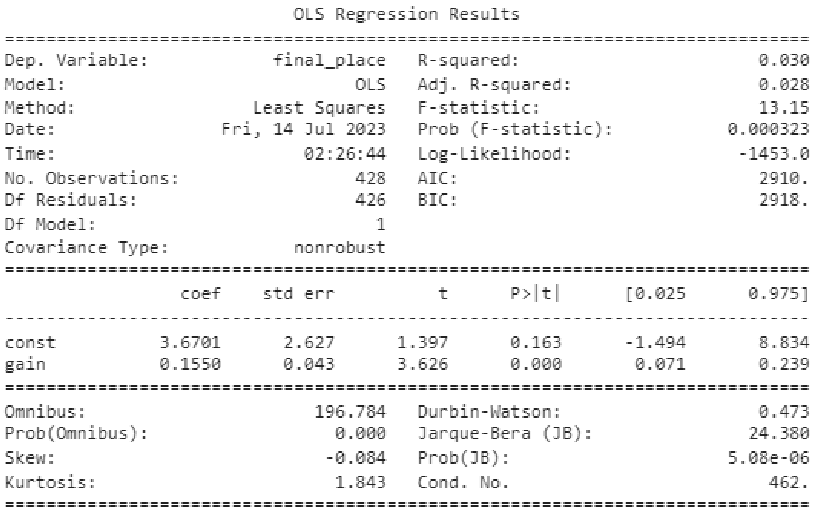
[11] Megginson, T. (2022, May 19). *Perspective | eurovision has always been a forum for political performance*. The Washington Post. https://www.washingtonpost.com/outlook/2022/05/19/eurovision-has-always-been-forum-political-performance/

[12] Ray, M. (2023, June 20). *Eurovision Song Contest*. Encyclopædia Britannica. https://www.britannica.com/art/Eurovision-Song-Contest

[13] Tollefson, J. (2022). *Climate change is hitting the planet faster than scientists originally thought*. Nature. https://doi.org/10.1038/d41586-022-00585-7

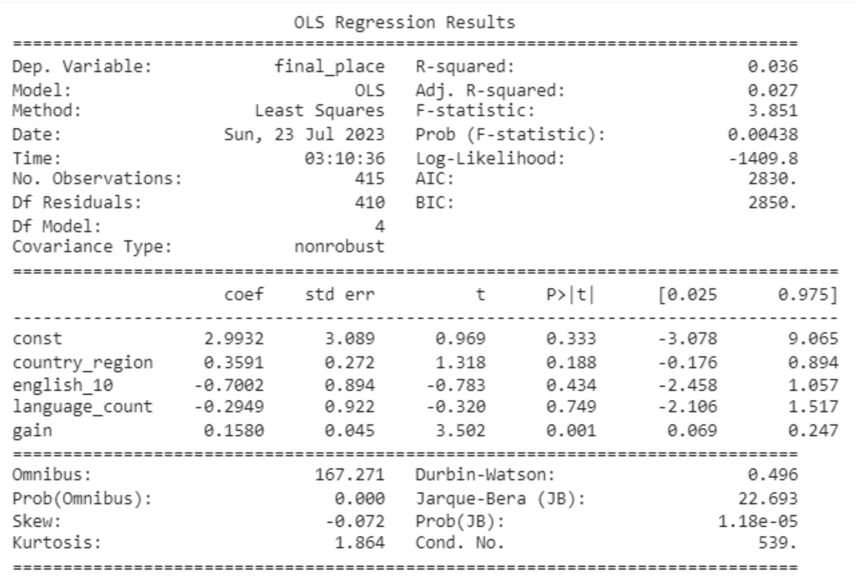
[14] Tsioulcas, A. (2020, March 18). *Eurovision 2020 is canceled because of coronavirus*. NPR. https://www.npr.org/sections/coronavirus-live-updates/2020/03/18/817944536/eurovision-2020-is-canceled-because-of-coronavirus

8 Appendix



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.030
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     13.15
Date:                Fri, 14 Jul 2023   Prob (F-statistic):           0.000323
Time:                        02:26:44   Log-Likelihood:                -1453.0
No. Observations:                 428   AIC:                             2910.
Df Residuals:                     426   BIC:                             2918.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.6701      2.627      1.397      0.163      -1.494       8.834
gain           0.1550      0.043      3.626      0.000       0.071       0.239
==============================================================================
Omnibus:                      196.784   Durbin-Watson:                   0.473
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.380
Skew:                          -0.084   Prob(JB):                     5.08e-06
Kurtosis:                       1.843   Cond. No.                         462.
==============================================================================
```

Figure 28: Simple Linear Regression Model Summary (ND-GAIN Score)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     3.851
Date:                Sun, 23 Jul 2023   Prob (F-statistic):            0.00438
Time:                        03:10:36   Log-Likelihood:                -1409.8
No. Observations:                 415   AIC:                             2830.
Df Residuals:                     410   BIC:                             2850.
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              2.9932      3.089      0.969      0.333      -3.078       9.065
country_region     0.3591      0.272      1.318      0.188      -0.176       0.894
english_10        -0.7002      0.894     -0.783      0.434      -2.458       1.057
language_count    -0.2949      0.922     -0.320      0.749      -2.106       1.517
gain               0.1580      0.045      3.502      0.001       0.069       0.247
==============================================================================
Omnibus:                      167.271   Durbin-Watson:                   0.496
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.693
Skew:                          -0.072   Prob(JB):                     1.18e-05
Kurtosis:                       1.864   Cond. No.                         539.
==============================================================================
```

Figure 29: Multiple Linear Regression Model Summary (ND-GAIN Score)

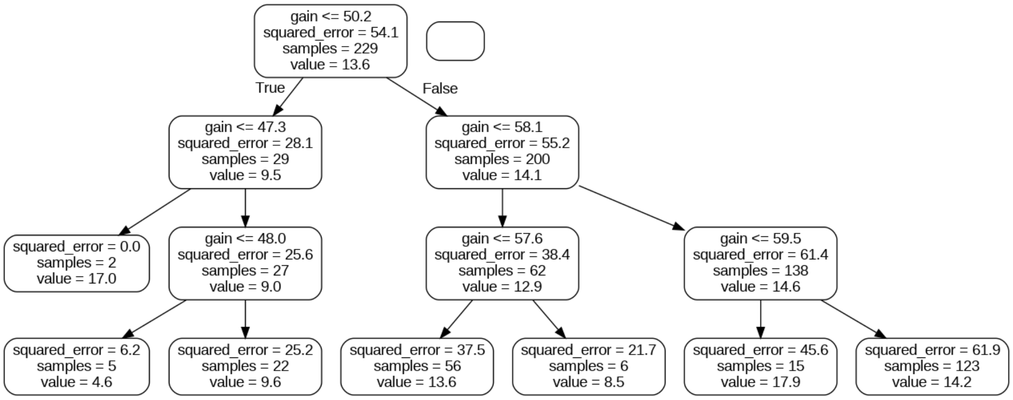


Figure 30: Random Forest Regression Tree 1 (ND-GAIN Score)

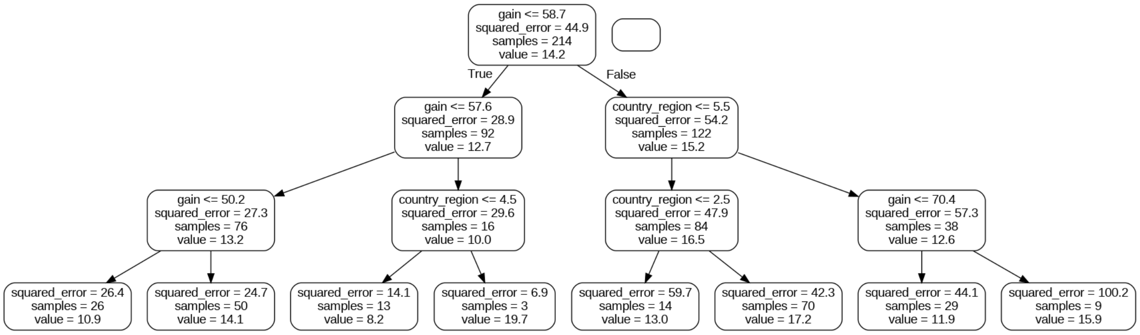
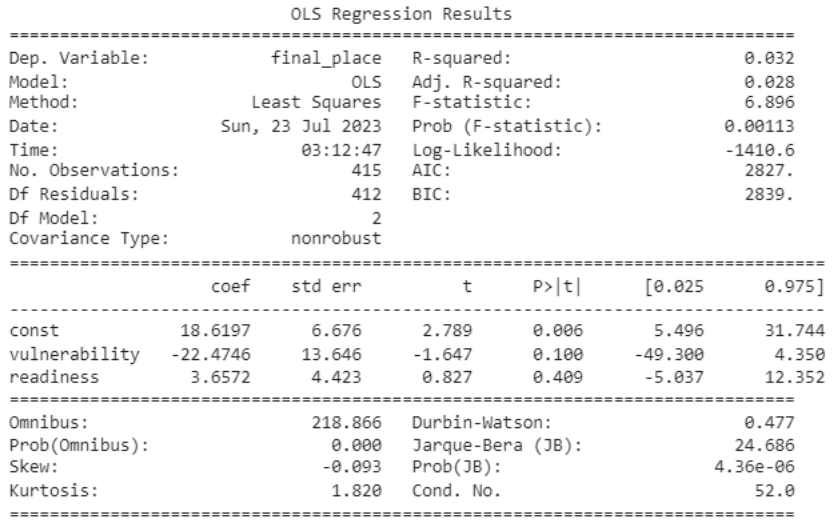
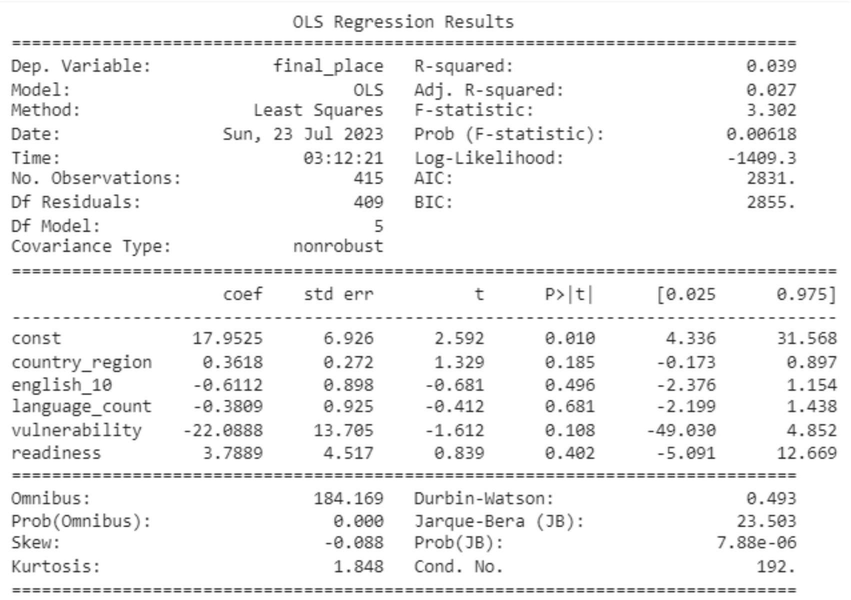


Figure 31: Random Forest Regression Tree 2 (ND-GAIN Score)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.032
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     6.896
Date:                Sun, 23 Jul 2023   Prob (F-statistic):            0.00113
Time:                        03:12:47   Log-Likelihood:                -1410.6
No. Observations:                 415   AIC:                             2827.
Df Residuals:                     412   BIC:                             2839.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            18.6197      6.676      2.789      0.006       5.496      31.744
vulnerability   -22.4746     13.646     -1.647      0.100     -49.300       4.350
readiness         3.6572      4.423      0.827      0.409      -5.037      12.352
==============================================================================
Omnibus:                      218.866   Durbin-Watson:                   0.477
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.686
Skew:                          -0.093   Prob(JB):                     4.36e-06
Kurtosis:                       1.820   Cond. No.                         52.0
==============================================================================
```

Figure 32: Multiple Linear Regression Model 1 Summary (ND-GAIN Indicators)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.039
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     3.302
Date:                Sun, 23 Jul 2023   Prob (F-statistic):            0.00618
Time:                        03:12:21   Log-Likelihood:                -1409.3
No. Observations:                 415   AIC:                             2831.
Df Residuals:                     409   BIC:                             2855.
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             17.9525      6.926      2.592      0.010       4.336      31.568
country_region     0.3618      0.272      1.329      0.185      -0.173       0.897
english_10        -0.6112      0.898     -0.681      0.496      -2.376       1.154
language_count    -0.3809      0.925     -0.412      0.681      -2.199       1.438
vulnerability    -22.0888     13.705     -1.612      0.108     -49.030       4.852
readiness          3.7889      4.517      0.839      0.402      -5.091      12.669
==============================================================================
Omnibus:                      184.169   Durbin-Watson:                   0.493
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.503
Skew:                          -0.088   Prob(JB):                     7.88e-06
Kurtosis:                       1.848   Cond. No.                         192.
==============================================================================
```

Figure 33: Multiple Linear Regression Model 2 Summary (ND-GAIN Indicators)

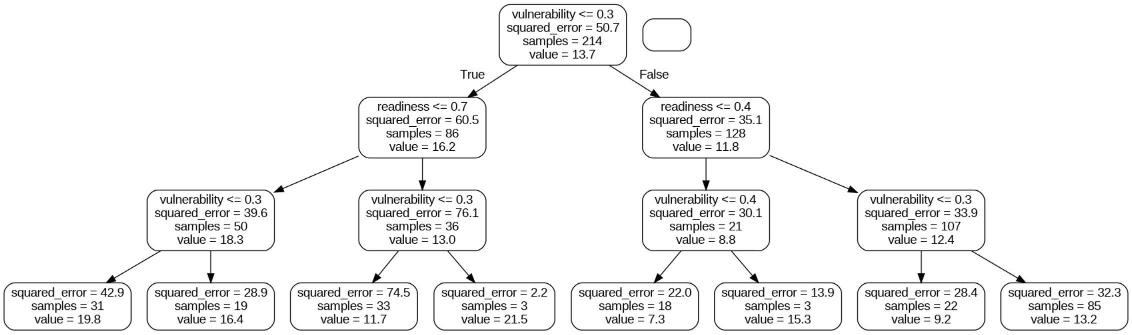


Figure 34: Random Forest Regression Tree 1 (ND-GAIN Indicators)

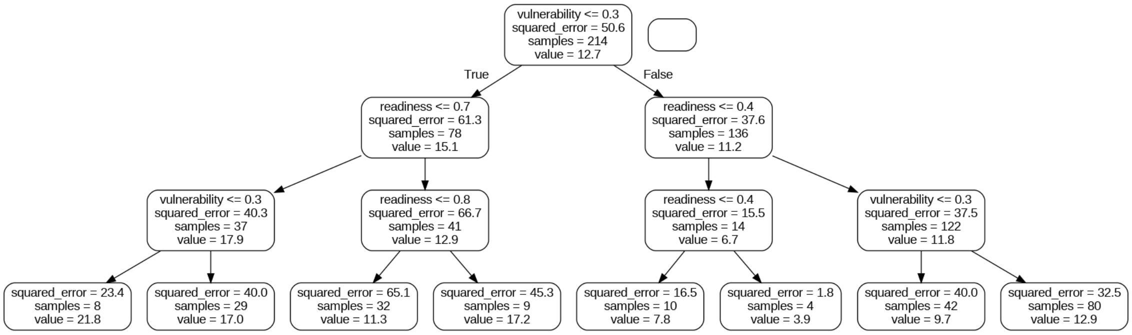
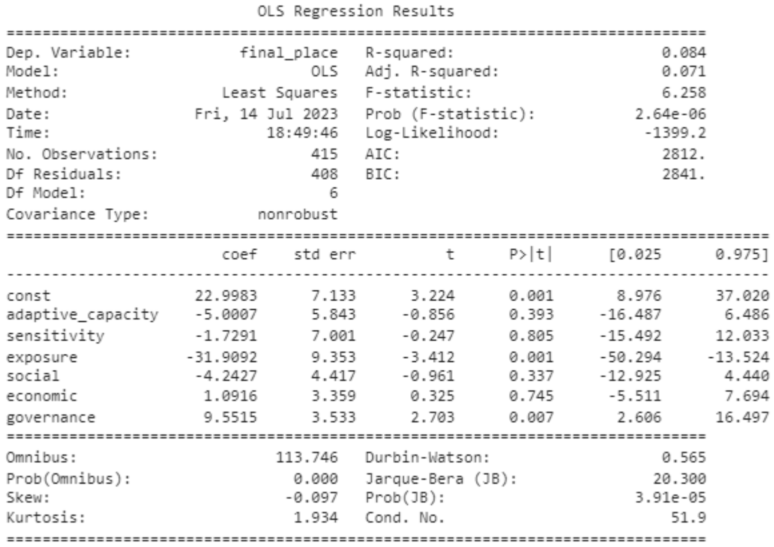
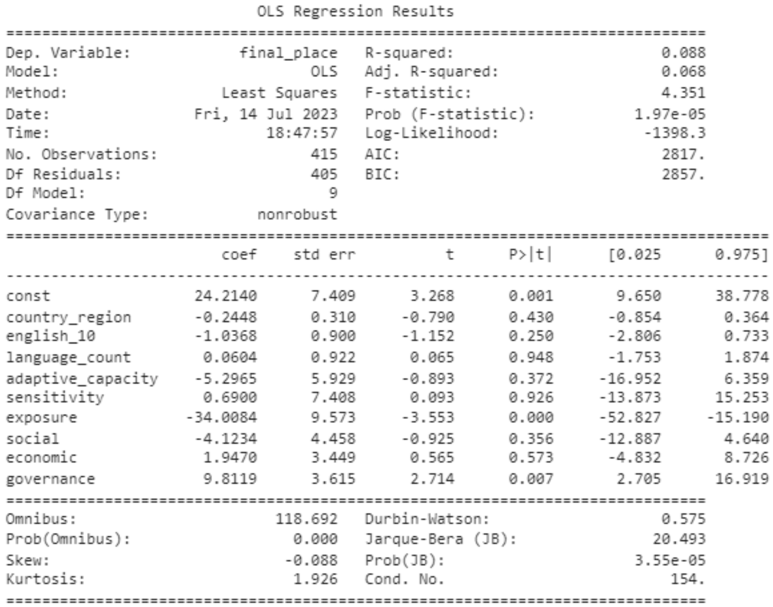


Figure 35: Random Forest Regression Tree 2 (ND-GAIN Indicators)



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.084
Model:                            OLS   Adj. R-squared:                  0.071
Method:                 Least Squares   F-statistic:                     6.258
Date:                Fri, 14 Jul 2023   Prob (F-statistic):           2.64e-06
Time:                        18:49:46   Log-Likelihood:                -1399.2
No. Observations:                 415   AIC:                             2812.
Df Residuals:                     408   BIC:                             2841.
Df Model:                           6                                         
Covariance Type:            nonrobust                                         
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                22.9983      7.133      3.224      0.001       8.976      37.020
adaptive_capacity    -5.0007      5.843     -0.856      0.393     -16.487       6.486
sensitivity          -1.7291      7.001     -0.247      0.805     -15.492      12.033
exposure            -31.9092      9.353     -3.412      0.001     -50.294     -13.524
social               -4.2427      4.417     -0.961      0.337     -12.925       4.440
economic              1.0916      3.359      0.325      0.745      -5.511       7.694
governance            9.5515      3.533      2.703      0.007       2.606      16.497
==============================================================================
Omnibus:                      113.746   Durbin-Watson:                   0.565
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.300
Skew:                          -0.097   Prob(JB):                     3.91e-05
Kurtosis:                       1.934   Cond. No.                         51.9
==============================================================================
```

Figure 36: Multiple Linear Regression Model 1 Summary (ND-GAIN Indicator Components)



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            final_place   R-squared:                       0.088
Model:                            OLS   Adj. R-squared:                  0.068
Method:                 Least Squares   F-statistic:                     4.351
Date:                Fri, 14 Jul 2023   Prob (F-statistic):           1.97e-05
Time:                        18:47:57   Log-Likelihood:                -1398.3
No. Observations:                 415   AIC:                             2817.
Df Residuals:                     405   BIC:                             2857.
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                24.2140      7.409      3.268      0.001       9.650      38.778
country_region       -0.2448      0.310     -0.790      0.430      -0.854       0.364
english_10           -1.0368      0.900     -1.152      0.250      -2.806       0.733
language_count        0.0604      0.922      0.065      0.948      -1.753       1.874
adaptive_capacity    -5.2965      5.929     -0.893      0.372     -16.952       6.359
sensitivity           0.6900      7.408      0.093      0.926     -13.873      15.253
exposure            -34.0084      9.573     -3.553      0.000     -52.827     -15.190
social               -4.1234      4.458     -0.925      0.356     -12.887       4.640
economic              1.9470      3.449      0.565      0.573      -4.832       8.726
governance            9.8119      3.615      2.714      0.007       2.705      16.919
==============================================================================
Omnibus:                      118.692   Durbin-Watson:                   0.575
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.493
Skew:                          -0.088   Prob(JB):                     3.55e-05
Kurtosis:                       1.926   Cond. No.                         154.
==============================================================================
```

Figure 37: Multiple Linear Regression Model 2 Summary (ND-GAIN Indicator Components)

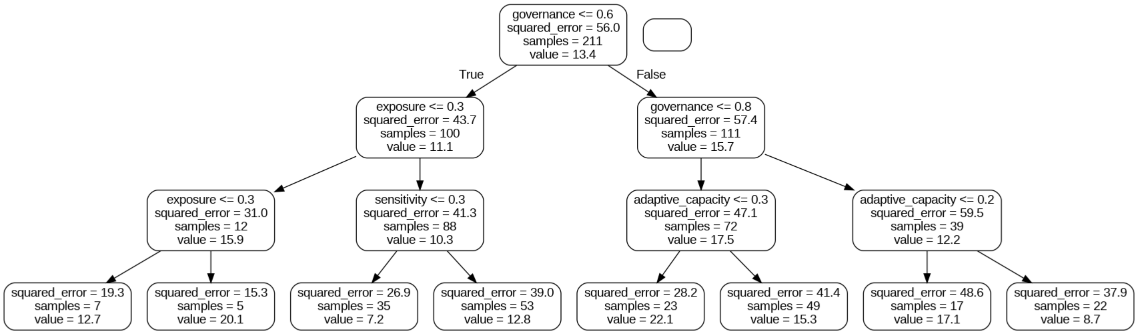


Figure 38: Random Forest Regression Tree 1 (ND-GAIN Indicator Components)

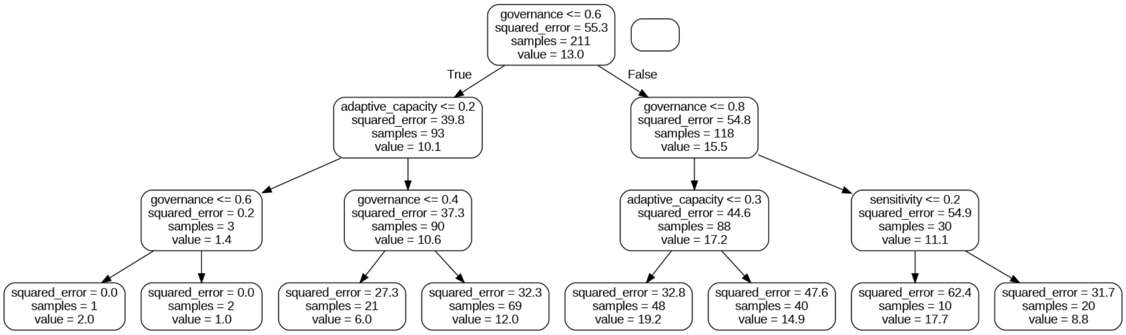
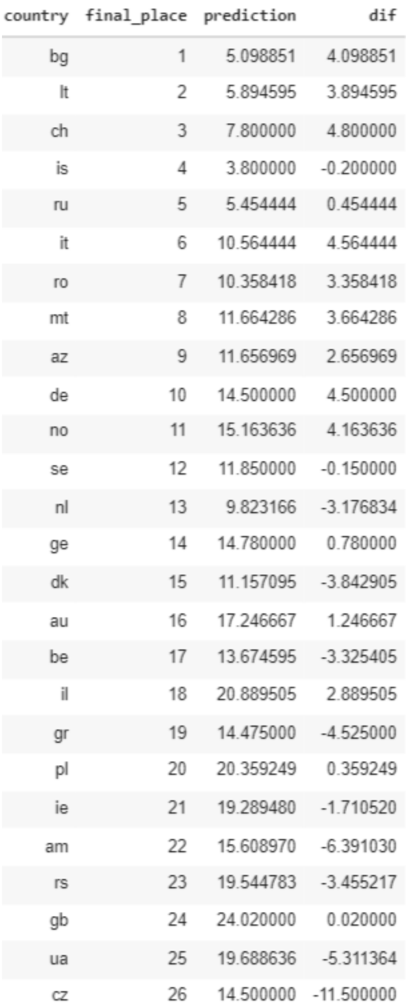


Figure 39: Random Forest Regression Tree 2 (ND-GAIN Indicator Components)



country	final_place	prediction	dif
bg	1	5.098851	4.098851
lt	2	5.894595	3.894595
ch	3	7.800000	4.800000
is	4	3.800000	-0.200000
ru	5	5.454444	0.454444
it	6	10.564444	4.564444
ro	7	10.358418	3.358418
mt	8	11.664286	3.664286
az	9	11.656969	2.656969
de	10	14.500000	4.500000
no	11	15.163636	4.163636
se	12	11.850000	-0.150000
nl	13	9.823166	-3.176834
ge	14	14.780000	0.780000
dk	15	11.157095	-3.842905
au	16	17.246667	1.246667
be	17	13.674595	-3.325405
il	18	20.889505	2.889505
gr	19	14.475000	-4.525000
pl	20	20.359249	0.359249
ie	21	19.289480	-1.710520
am	22	15.608970	-6.391030
rs	23	19.544783	-3.455217
gb	24	24.020000	0.020000
ua	25	19.688636	-5.311364
cz	26	14.500000	-11.500000

Figure 40: 2020 Prediction Comparison