Classification of Regions Based on Countries Alcohol Consumption

Team Good Spirits:
Adam Rockett, Caelin Schaefer,
Fiona Cleary, Elias Peters

Abstract: Our objective of this project is to analyze and classify countries into regions based on consumption of different types of alcohol. The research and results utilizes random forests, and CART programming to properly classify variables into the appropriate categories. We found that, since our data set had such a small amount of observations from the Pacific Islands and the Caribbean, both of the algorithms had a tough time classifying countries into those regions. On the other hand, the algorithms were successful at classifying countries into Africa, Asia and Europe. It was able to correctly classify countries in these regions more than half of the time, and sometimes with accuracies up to 77.8%.

Background and Significance: Alcoholic beverages are consumed by people across the globe. Our goal is to compare intake of alcohol throughout various regions of the world. We classified 193 different countries into six geographical regions, and evaluated different types of alcohol consumption for each country. The data gave us information about the number of beer, wine, and spirits servings, as well as total alcohol consumption in liters per person for each country. Knowing how much alcohol people intake is valuable data to have, and being allows us able to detect global patterns in alcohol consumption could be utilized in a number of ways. Doctors and insurance companies have an interest in knowing which populations will be more exposed to alcohol related illnesses; in 2016, more than three million people died as a result of alcohol consumption, and these deaths accounted for around 5% of all deaths that year.[1] Additionally, the market for alcohol is very profitable with the total export of liquor valued at $6.05 billion USD[2] in 2016. Distilleries, breweries, and vineyards could use information about different regional markets to become more profitable.

Methods: We obtained our data from the website FiveThirtyEight. Our data was collected by Mona Chalabi, who gathered and analyzed information from the World Health Organization about different kinds of alcohol consumption across the globe. Chalabi's intention was to figure out which countries drank the most of different kinds types of alcohol, and also which countries mostly abstained from drinking. We sorted each of the different countries into geographic regions, then ran random forest simulations and CART analysis.

When classifying the countries into regions, we initially divided the globe into nine different regions: Africa, Asia, the Caribbean, Central America, Europe, the Middle East, North America, the Pacific Islands, and South America. We put Australia in the Pacific Islands group so that it would not be in its own group. After running our data with these groupings, it became apparent that the data was not running well due to the small size of some of the geographic regions. As a result, we reclassified our data into six regions: Africa, the Americas, Asia, the Caribbean, Europe, and the Pacific Islands. We sorted the countries in the Middle East into Asia because there were so many countries where there was no reported alcohol consumption, making it difficult for our algorithms to classify countries in the region effectively. We lumped all of the countries in North, Central, and South America together because each of those regions contained a fairly small number of countries as well. We decided to keep the Caribbean separate from the Americas because there were a larger number of countries within that region, and we decided to keep the Pacific Islands separate from Asia for the same reason.

Results: When we ran the random forest algorithm on the data, we got an OOB estimate of error rate of 43.75% and produced the confusion matrix below in Figure 1. We ended up running the random forest with 100 trees and with 2 variables selected randomly at each split since it is the combination that gave the lowest OOB error rate. When we tried changing these numbers, the OOB error rate tended to stay between 45% and 55%, even if the number of trees was increased to way larger numbers such as 10,000.

As can be seen from Figure 1, our random forest can predict the countries in Africa, the Americas, Asia and Europe from the training set fairly well, but it fails to predict any of the countries from the Caribbean and Pacific Islands. This could be because there was a relatively small amount of data points from those regions in the training set.

1 "Global Status Report on Alcohol and Health 2018." *World Health Organization*, World Health Organization, 21 Sept. 2018.

2 "Liquor Global Export and Top Exporting Countries." *Tridge*.

```
        OOB estimate of  error rate: 43.75%
Confusion matrix:
                Africa America Asia Carribean Europe Pacific Islands class.error
Africa              17       1    9         0      3                0   0.4333333
America              1       5    1         1      1                0   0.4444444
Asia                 5       1   17         0      1                0   0.2916667
Carribean            1       2    0         2      1                0   0.6666667
Europe               2       0    4         1     13                0   0.3500000
Pacific Islands      3       1    2         0      1                0   1.0000000
```

Figure 1: Confusion Matrix of Training Data for Random Forest.

	observed					
predicted	Africa	America	Asia	Carribean	Europe	Pacific Islands
Africa	18	1	5	0	1	3
America	1	6	2	2	0	1
Asia	8	0	11	2	2	2
Carribean	0	0	0	2	4	0
Europe	1	5	0	0	18	2
Pacific Islands	0	0	0	0	0	0

Figure 2: Table of Prediction Set Ran Through Random Forest

We also used the same training and verification data sets using the CART classification algorithm to see if it could more accurately categorize our data. The tree that the algorithm eventually fit to our data is given in Figure 3. Right away, a problem that we can see is that no matter what values a country has for the different variables, it will never get classified as being from the Caribbean or the Pacific Islands. This is consistent with our analysis of the random forest, and could just be a shortcoming of the data we had. When we use this tree to classify the prediction data set, we get the following results in Figure 4.

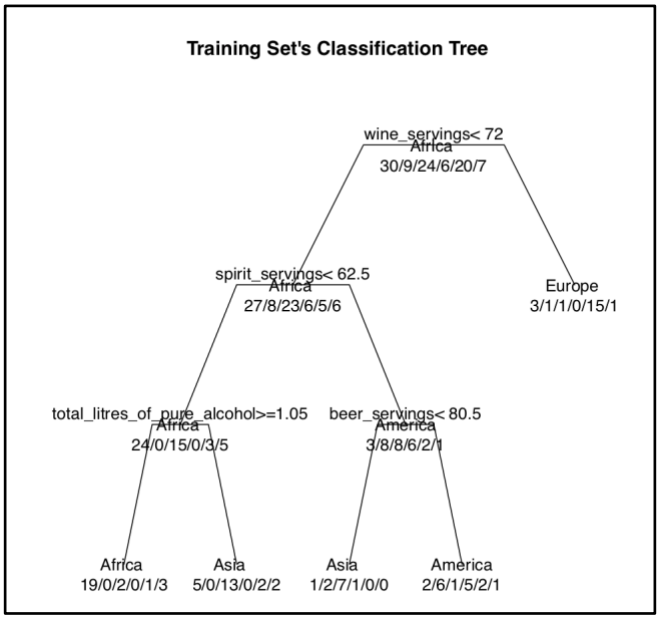

Figure 3: CART Tree for our data set.

predictions	Africa	America	Asia	Carribean	Europe	Pacific Islands
Africa	15	1	3	0	0	2
America	1	4	1	5	6	1
Asia	11	2	14	1	2	3
Carribean	0	0	0	0	0	0
Europe	1	5	0	0	17	2
Pacific Islands	0	0	0	0	0	0

Figure 4: Table of Prediction Set Using CART Tree

Discussion/Conclusions: As with all research, there are limitations on what can be learned from our results. In all regions there are vast differences between different countries which create outliers. Our analysis did not take into account variations in alcohol legislation between countries, and did not evaluate cultural and attitude differences towards alcohol. Furthermore, the accuracy of our conclusions are closely related to what regions are represented in our training dataset. Random forests require that the training data be randomly selected, however, over or under representation of different regions in our training data increases error in our results.

From Figure 2, when we ran the other half of the data set through the random forest to verify if it was a good classifier of the data, we could see that it does a decent job of predicting the region if the country was from Africa, Asia and Europe, where it accurately predicted the countries 64.3%, 57.9% and 72.0% respectively. However, we can see that our random forest never predicted a country was from the Pacific Islands, and is below average at predicting a country is from the Caribbean.

From Figure 4, we can see that our model for predicting the data using CART does not to a very good job of classifying most of the regions. The only two regions where it classifies the majority of the countries in the correct region is Asia and Europe, where it predicted 77.8% and 68.0% of the countries respectively. The only other region that gets above 50% of the countries correct is Africa, but it only predicts 53.6% correct. All of the other regions have prediction accuracies below 50%.

The Pacific Islands were an interesting case where we were unable to predict that any country would actually belong to the Pacific Islands, however, as is apparent from Figure 2, we predicted that Pacific Island countries belonged to a diverse group of regions. What's particularly interesting is that our CART analysis also did not classify any countries into the Pacific Islands, and instead classified countries from the Pacific Islands into every region except the Pacific Islands and the Caribbean (see Figure 4). There are several possible explanations for this. It could mean that there is significant similarity in alcoholic consumption between different countries in the Pacific Islands and another global region, possibly due to distinctive cultural differences between countries. Another explanation could be that the Pacific Islands should have been grouped into another region, because they were one of our smallest samples in our training data. We have particularly high error rates from both CART and random forest for both the Pacific Islands and the Caribbean, which had 15 countries and 12 countries respectively. The other regions had 21-58 countries to sample from and thus could build a better classifications more quickly.

References:

"A Very Basic Introduction to Random Forests Using R." *Oxford Protein Informatics Group*.

Bhalla, Deepanshu. "Simple Introduction to Random Forest Using R." *ListenData*.

Chalabi, Mona. "Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?" *FiveThirtyEight*, FiveThirtyEight, 13 Aug. 2014.

"Global Status Report on Alcohol and Health 2018." *World Health Organization*, World Health Organization, 21 Sept. 2018.

"Liquor Global Export and Top Exporting Countries." *Tridge*.

Srivastava, Tavish. "Tuning Random Forest Model | Machine Learning | Predictive Modeling." *Analytics Vidhya*, 26 Apr. 2018.

Appendix:

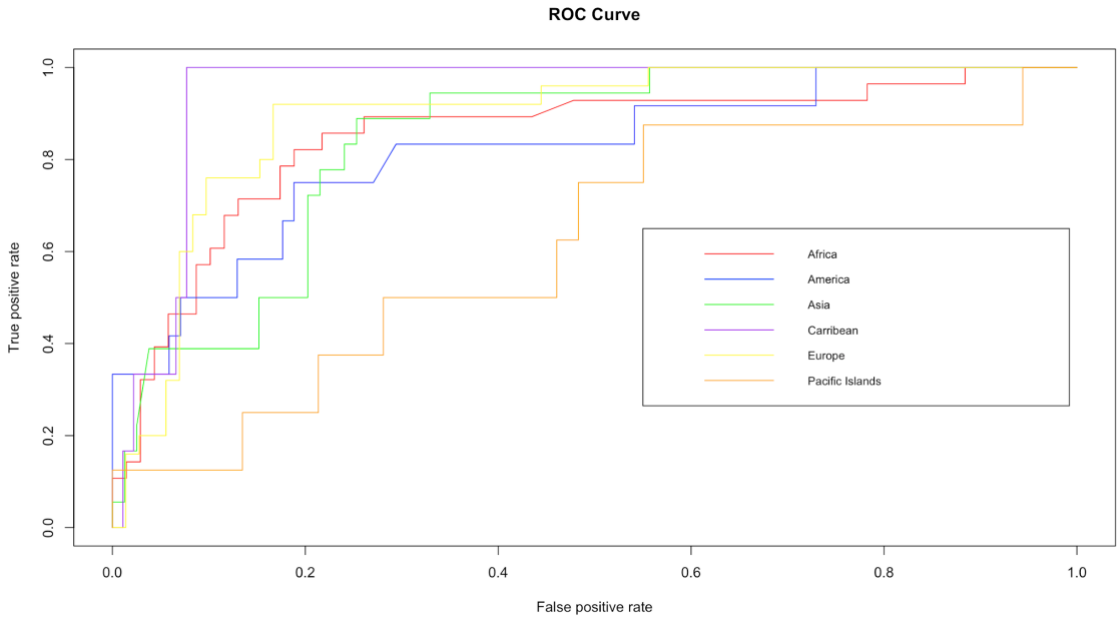

Figure 5: ROC Curves From Random Forest

The ROC curves in Figure 5 also support that our variables aren't a very good predictor of Pacific Island countries since the ROC curve for that specific region is very close to the 45 degree line.

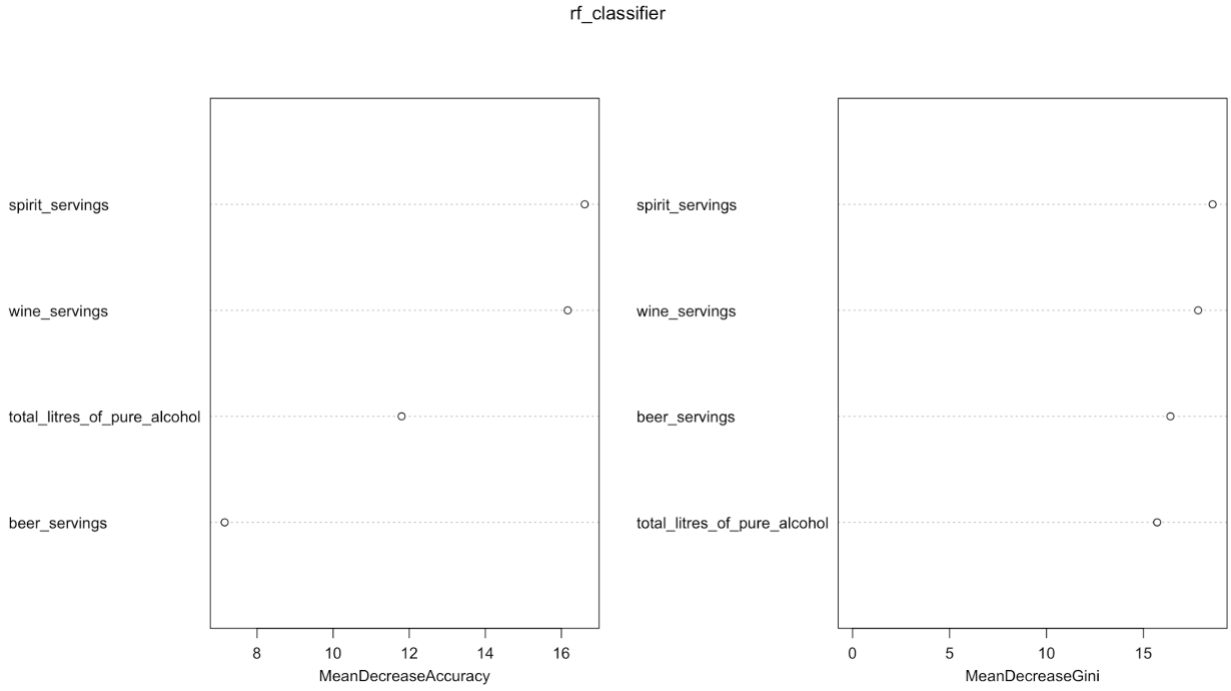

Figure 6: Importance Plot From Random Forest

The importance plot in Figure 6 shows that for both accuracy and the Gini coefficient measures, the two most important variables are the number of spirits served per person and the number of

wine servings per person. This is a good indication that our random forest is as accurate as possible given the data.