EFFECTS OF AGE, LOCATION AND ETHNICITY ON CHRONIC DISEASES

KARMISHTHA SETH

JULY 18, 2020

1. INTRODUCTION

The CDC classifies Chronic Diseases as "...conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both", common examples being cancer, heart disease, diabetes and asthma. Common causes for such chronic diseases are tobacco usage, excessive alcohol intake, poor nutrition etc. Chronic diseases impact almost every human being's life, and in a major way. Studies have shown that chronic diseases are quite biased in terms of the groups that are targeted by them. A study done in the UK, testing effects of race on diabetes and chronic kidney disease showed that there was a significant increase in the aforementioned diseases in the Black and South Asian community when compared to Caucasians[1]. Another study performed by Italian General Practitioners looked at how ageing had an effect on chronic disease, and noted that there is a significant increase in the prevalence of chronic diseases after the age of 60[2]. Looking at such data, I was motivated to look at what effect age, ethnicity and location have on chronic disease prevalence.

2. METHODS

I collected my initial data set for chronic diseases was a CDC dataset collected from Kaggle, labelled "U.S._Chronic_Disease_Indicators"[3] (Figure 1.) from which I used the variables "LocationAbbr", "Stratification1", "Topic", "LocationDesc", "DataValue".

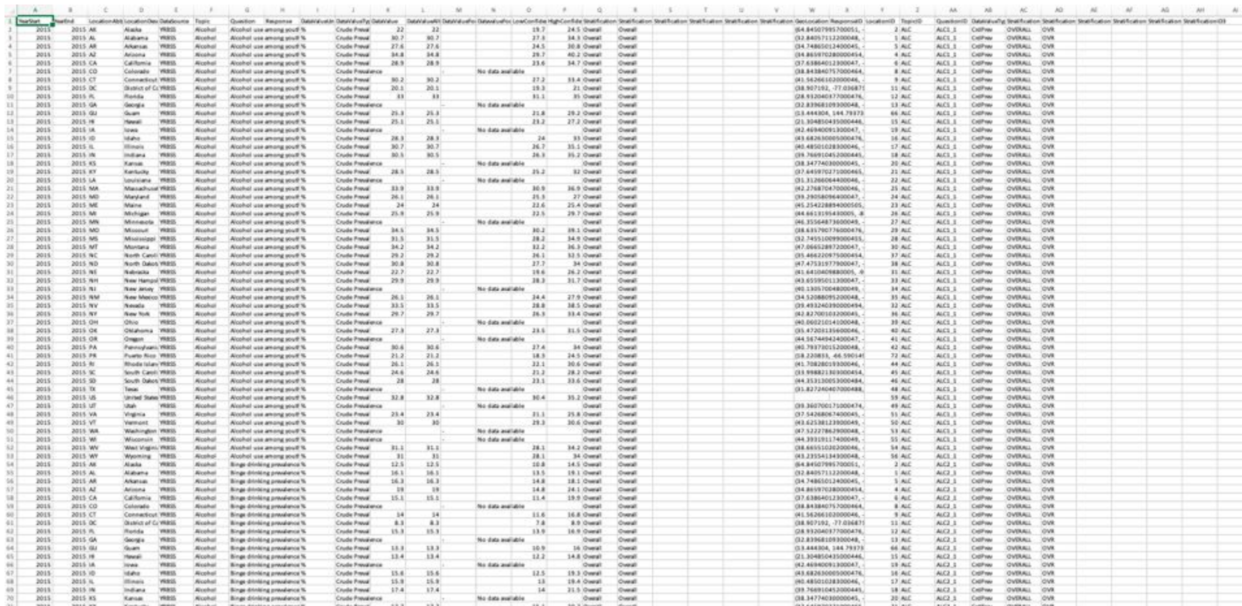

Figure 1. U.S._Chronic_Disease_Indicators dataset

The U.S._Chronic_Diseases_Indicators dataset consists of 403985 rows of data and 34 variable columns. Out of the 34 variables, I used 3 of them, specifically "Stratification1", "Topic", and "LocationDesc" (FIGURE 2).

	LocationDesc	Topic	Stratification1
1	Alaska	Alcohol	Overall
2	Alabama	Alcohol	Overall
3	Arkansas	Alcohol	Overall
4	Arizona	Alcohol	Overall
5	California	Alcohol	Overall
6	Colorado	Alcohol	Overall
7	Connecticut	Alcohol	Overall
8	District of Columbia	Alcohol	Overall
9	Florida	Alcohol	Overall
10	Georgia	Alcohol	Overall
11	Guam	Alcohol	Overall
12	Hawaii	Alcohol	Overall
13	Iowa	Alcohol	Overall
14	Idaho	Alcohol	Overall
15	Illinois	Alcohol	Overall
16	Indiana	Alcohol	Overall
17	Kansas	Alcohol	Overall

Figure 2. Required columns from dataset

To analyse my question, how likely is someone belonging to a certain ethnicity and location to have a chronic disease, I decided to use a decision tree. Before creating my decision tree, random forest results and predictions, I looked at the distributions of the different variables and their density plots to look at how much overlap there was. Decision trees help lay the problem out in a very simple manner, and allow us to fully analyse the consequences of any decision. It provides a framework to quantify the values of outcomes and the probabilities of achieving them. For my decision tree, I used Topic as my dependent variable, and Stratification and Location as my predictor variables. I then performed Random Forest on my data. I used random forest because since it consists of multiple decision trees, it provides higher accuracy, especially with larger data[4]. Since all my variables of interests were categorical variables, I factored them out, where each category is represented by a certain number (Figure 3). Based on my random forest results, I created predictions for how likely someone is to have a chronic disease based on their location and stratification.

```
> DI
   Topic LocationDesc Stratification1
1      1            1               1
2      1            2               1
3      1            3               1
4      1            4               1
5      1            5               1
6      1            6               1
7      1            7               1
8      1            8               1
9      1            9               1
10     1           10               1
11     1           11               1
12     1           12               1
```

Figure 3. Factored variables

3. RESULTS

My distribution charts showed a higher prevalence of chronic diseases in more minority ethnic groups than caucasians (Figure 4) and I expected cardiac diseases and diabetes to be the most common chronic diseases, which is also shown (Figure 5). Since I have such large data, I performed hexagonal binning to look at how much overlap there is between stratification and location, and as shown in Figure 6, there is significant overlap between the two. Based on my decision tree I created (Figure 7), by just looking at it, I am able to predict which disease one is more likely to have based on their stratification and location. However, I did encounter an issue, where my decision tree is extremely condensed, which is primarily due to the large number of categories in each variable. On performing my random forest (Figure 8), I did notice a large OOB estimate of error rate, which could be attributed to the large number of categorical data. To better understand my random forest results, I ran a prediction and confusion matrix, examples of which can be found in Figure 9.

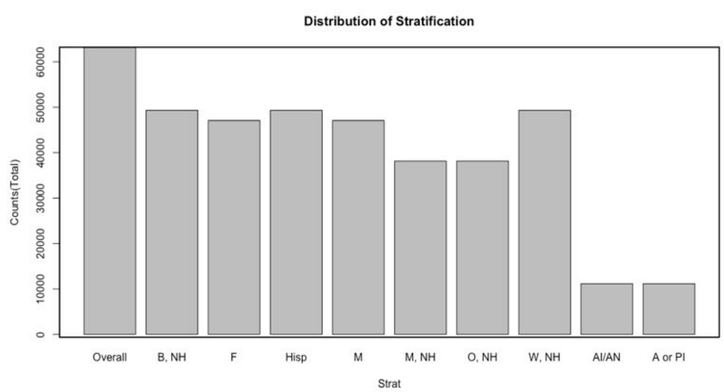

Figure 4. Distribution of Stratification

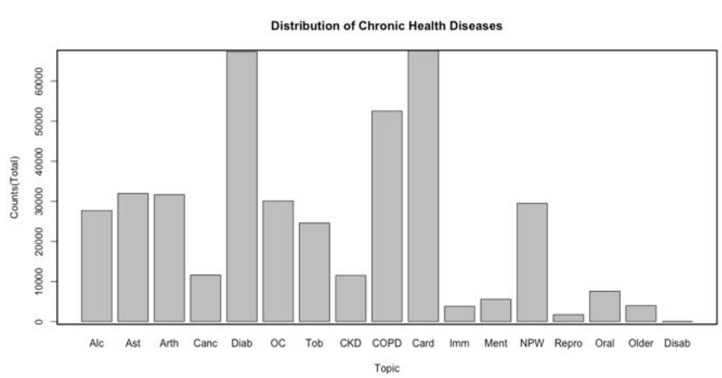

Figure 5. Distribution of Chronic Health Diseases

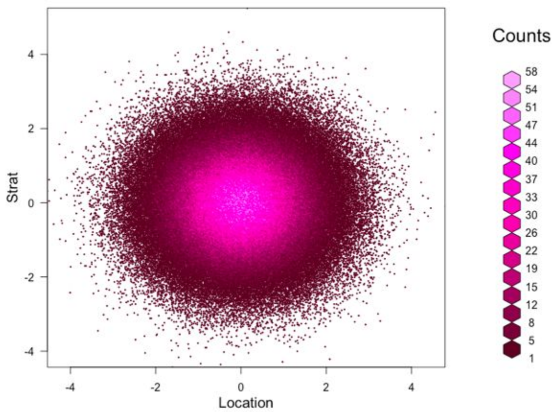

Figure 6. Hexagonal Binning showing Overlap between Location and Stratification.

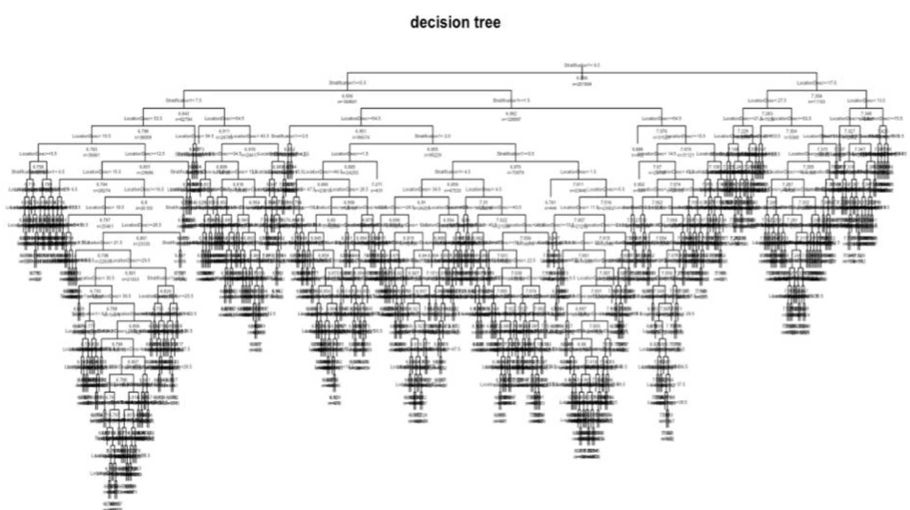

Figure 7. Decision Tree

```
Call:
 randomForest(formula = Topic ~ Stratification1, data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 82.22%
Confusion matrix:
                                               Alcohol Arthritis Asthma Cancer Cardiovascular Disease Chronic Kidney Disease Chronic Obstructive Pulmonary Disease Diabetes Disability
Alcohol                                              0         0      0      0                    784                      0                                     0    13064          0
Arthritis                                            0         0      0      0                      0                      0                                     0    15826          0
Asthma                                               0         0      0      0                    985                      0                                     0    14994          0
Cancer                                               0         0      0      0                      0                      0                                     0     5802          0
Cardiovascular Disease                               0         0      0      0                   4079                      0                                     0    27260          0
Chronic Kidney Disease                               0         0      0      0                    792                      0                                     0     4970          0
Chronic Obstructive Pulmonary Disease                0         0      0      0                   1984                      0                                     0    24284          0
Diabetes                                             0         0      0      0                   1828                      0                                     0    31843          0
Disability                                           0         0      0      0                      0                      0                                     0       28          0
Immunization                                         0         0      0      0                      0                      0                                     0     1919          0
Mental Health                                        0         0      0      0                      0                      0                                     0     2798          0
Nutrition, Physical Activity, and Weight Status      0         0      0      0                      0                      0                                     0    14756          0
Older Adults                                         0         0      0      0                      0                      0                                     0     2002          0
Oral Health                                          0         0      0      0                      0                      0                                     0     3796          0
Overarching Conditions                               0         0      0      0                    764                      0                                     0    14282          0
Reproductive Health                                  0         0      0      0                      0                      0                                     0      868          0
```

Figure 8. Sample of Random Forest Results

```
> pred
       2        3        5        6        7        8       14       18       20       22       24       27 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      29       32       33       34       35       36       38       41       42       46       47       49 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      50       52       54       56       57       59       61       63       64       65       70       72 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      73       74       75       78       79       80       86       87       88       89       94       96 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      98       99      101      103      106      108      109      115      117      122      123      124 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     126      127      128      136      137      139      142      150      152      153      156      157 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     158      160      162      163      166      167      169      170      173      174      175      176 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     178      179      180      186      189      191      196      199      200      201      202      207 
```

```
> cm
                      pred
                       Alcohol Arthritis Asthma Cancer Cardiovascular Disease Chronic Kidney Disease
  Alabama                    0         0      0      0                    200                      0
  Alaska                     0         0      0      0                    195                      0
  Arizona                    0         0      0      0                    243                      0
  Arkansas                   0         0      0      0                    205                      0
  California                 0         0      0      0                    211                      0
  Colorado                   0         0      0      0                    218                      0
  Connecticut                0         0      0      0                    179                      0
  Delaware                   0         0      0      0                    203                      0
  District of Columbia       0         0      0      0                    188                      0
  Florida                    0         0      0      0                    266                      0
  Georgia                    0         0      0      0                    186                      0
  Guam                       0         0      0      0                      0                      0
  Hawaii                     0         0      0      0                    250                      0
  Idaho                      0         0      0      0                    184                      0
  Illinois                   0         0      0      0                    207                      0
  Indiana                    0         0      0      0                    199                      0
```

Figure 9. Prediction and Confusion Matrix

4. CONCLUSION

There is a significant interaction between the type of chronic disease, and location and stratification. Using my decision tree and prediction matrix, I am able to gauge what chronic disease someone of a certain stratification and location is most likely to have. I can apply this knowledge to analysing deaths due to COVID-19 and predict how likely someone is to die based on their age and ethnicity. My data and results supported my hypothesis of there being a greater prevalence of chronic diseases in minority ethnic groups when compared to white, non hispanic.

Figure 22. Death counts due to COVID-19 based on race

5. BIBLIOGRAPHY

1. Dreyer, G., et al. “Effect of Ethnicity on the Prevalence of Diabetes and Associated Chronic Kidney Disease.” *OUP Academic*, Oxford University Press, 15 Jan. 2009, academic.oup.com/qjmed/article/102/4/261/1550086.

2. Atella, Vincenzo, et al. “Trends in Age-Related Disease Burden and Healthcare Utilization.” *Aging Cell*, John Wiley and Sons Inc., Feb. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6351821/.

3. Centers for Disease Control and Prevention. “Chronic Disease Indicators.” *Kaggle*, 17 Aug. 2017, www.kaggle.com/cdc/chronic-disease.

4. Deng, Houtao. “Why Random Forests Outperform Decision Trees.” *Medium*, Towards Data Science, 12 Dec. 2018, towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5.