

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 Caucasians¹. 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 60². Looking at such data, I was motivated to look at what effect age, ethnicity and location have on chronic disease prevalence.

COVID-19 has taken over the entire world in a mere 6 months, has caused multiple countries to shut down, lock borders, caused the market to be lower than it has been in decades and has even accelerated the speed of recession. While the effects of it will be long lasting, I am very intrigued to see how this virus disproportionately affects minority communities, and older age-d people.

Dr Chaand Nagpaul, the British Medical Association (BMA) council chair and a general practitioner (GP) in north London, UK said, “However, it is a clear and consistent theme from the reports and what we know about those who have died so far, that a disproportionate number of those health-care workers who have tragically lost their lives are from BAME communities.”³

In addition COVID-19 shows an increased number of cases and a greater risk of severe disease

Figure 1. U.S. Chronic Disease Indicators dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data as of	Start week	End Week	State	Age group	Race and His	COVID-19 De	Total Deaths	Pneumonia C	Pneumonia a	Influenza De	Pneumonia, I	Footnote
2	07/08/20	02/01/20	07/04/20	United States	All Ages	Total Deaths	114741	1324958	125868	49623	6492	196538	
3	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	60862	974415	81724	24734	4303	121940	
4	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	26426	177674	20910	11912	969	36084	
5	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	888	8187	982	472	68	1466	
6	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	5629	38200	5170	2546	267	8411	
7	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	127	1606	150	74	8	211	
8	07/08/20	02/01/20	07/04/20	United States	All Ages	Non-Hispanic	282	5941	414	127	34	601	
9	07/08/20	02/01/20	07/04/20	United States	All Ages	Hispanic or L	19409	114570	15797	9298	811	26437	
10	07/08/20	02/01/20	07/04/20	United States	All Ages	Unknown	1118	4365	721	460	32	1388	
11	07/08/20	02/01/20	07/04/20	United States	Under 1 year	Non-Hispanic	3	2948	26	0	8	37	
12	07/08/20	02/01/20	07/04/20	United States	1-4 years	Non-Hispanic	3	592	18	1	18	38	
13	07/08/20	02/01/20	07/04/20	United States	5-14 years	Non-Hispanic	2	1007	33	1	26	60	
14	07/08/20	02/01/20	07/04/20	United States	15-24 years	Non-Hispanic	24	6148	105	7	30	151	
15	07/08/20	02/01/20	07/04/20	United States	25-34 years	Non-Hispanic	133	15196	392	63	68	530	
16	07/08/20	02/01/20	07/04/20	United States	35-44 years	Non-Hispanic	294	22213	812	103	115	1116	
17	07/08/20	02/01/20	07/04/20	United States	45-54 years	Non-Hispanic	1201	43659	2469	534	291	3420	
18	07/08/20	02/01/20	07/04/20	United States	55-64 years	Non-Hispanic	4412	112940	8311	1998	717	11402	
19	07/08/20	02/01/20	07/04/20	United States	65-74 years	Non-Hispanic	10409	187263	16364	4610	932	23043	
20	07/08/20	02/01/20	07/04/20	United States	75-84 years	Non-Hispanic	17140	248692	23444	7262	1050	34329	
21	07/08/20	02/01/20	07/04/20	United States	85 years and	Non-Hispanic	27241	333757	29750	10155	1048	47814	
22	07/08/20	02/01/20	07/04/20	United States	Under 1 year	Non-Hispanic	2	1975	19	1	3	23	
23	07/08/20	02/01/20	07/04/20	United States	1-4 years	Non-Hispanic	1	333	16	1	5	21	
24	07/08/20	02/01/20	07/04/20	United States	5-14 years	Non-Hispanic	7	457	14	0	10	31	
25	07/08/20	02/01/20	07/04/20	United States	15-24 years	Non-Hispanic	45	3238	56	12	5	92	
26	07/08/20	02/01/20	07/04/20	United States	25-34 years	Non-Hispanic	252	6027	252	107	37	428	
27	07/08/20	02/01/20	07/04/20	United States	35-44 years	Non-Hispanic	627	7990	565	272	46	962	
28	07/08/20	02/01/20	07/04/20	United States	45-54 years	Non-Hispanic	1376	14064	1411	780	133	2470	

Figure 2.

Deaths_involving_coronavirus_disease_2019__COVID-19__by_race_and_Hispanic_origin_group_and_age__by_state dataset

2 a. Chronic Diseases 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 3).

	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 3. 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 data⁷. Since all my variables of interests were categorical variables, I factored them out, where each category is represented by a certain number (Figure 4). 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 4. Factored variables

2 b. COVID-19 Dataset

The

Deaths_involving_coronavirus_disease_2019__COVID-19__by_race_and_Hispanic_origin_group_and_age__by_state.csv dataset consists of 4762 rows of data and 13 variable columns. Out of the 13 variables, I used 4 of them, specifically, “Age group”, “Race and Hispanic Origin Group”, “State” and “COVID-19 Deaths” (Figure 5).

	State	Age.group	Race.and.Hispanic.Origin.Group	COVID.19.Deaths
1	United States	All Ages	Total Deaths	114741
2	United States	All Ages	Non-Hispanic White	60862
3	United States	All Ages	Non-Hispanic Black	26426
4	United States	All Ages	Non-Hispanic American Indian or Alaska Native	888
5	United States	All Ages	Non-Hispanic Asian	5629
6	United States	All Ages	Non-Hispanic Native Hawaiian or Other Pacific Islander	127
7	United States	All Ages	Non-Hispanic More than one race	282
8	United States	All Ages	Hispanic or Latino	19409
9	United States	All Ages	Unknown	1118
10	United States	Under 1 year	Non-Hispanic White	3
11	United States	1-4 years	Non-Hispanic White	3
12	United States	5-14 years	Non-Hispanic White	2
13	United States	15-24 years	Non-Hispanic White	24
14	United States	25-34 years	Non-Hispanic White	133
15	United States	35-44 years	Non-Hispanic White	294
16	United States	45-54 years	Non-Hispanic White	1201
17	United States	55-64 years	Non-Hispanic White	4412

Figure 5. Required columns from dataset

To analyse my question, how likely is someone to die from COVID based on their age and ethnicity, I initially created distribution charts to see what percent of a certain age, location and ethnicity was represented in the data. Due to the large data size, I created a density plot, to look at the overlap of variables. Finally, to look at how age and ethnicity was related to the death count, I created a scatterplot comparing the two variables. I created a decision tree to help answer my question, then performed random forest, and based on the results of the random forest, created predictions for how likely someone of a certain age and ethnicity is to die from COVID.

3. RESULTS

3 a. Chronic Diseases Dataset

My distribution charts showed a higher prevalence of chronic diseases in more minority ethnic groups than caucasians (Figure 5) and I expected cardiac diseases and diabetes to be the most common chronic diseases, which is also shown (Figure 6). 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 7, there is significant overlap between the two. Based on my decision tree I created (Figure 8), 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 9), 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 10.

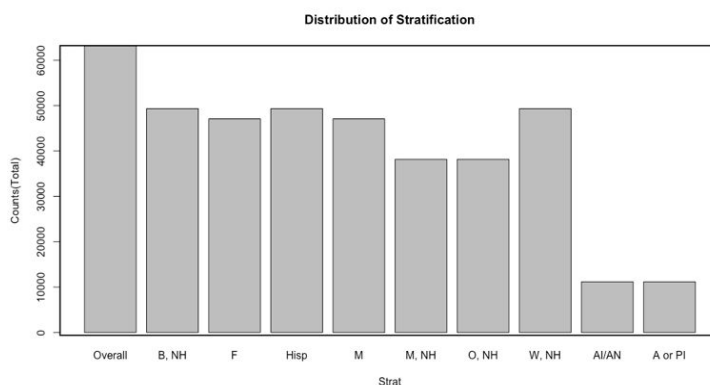


Figure 5. Distribution of Stratification

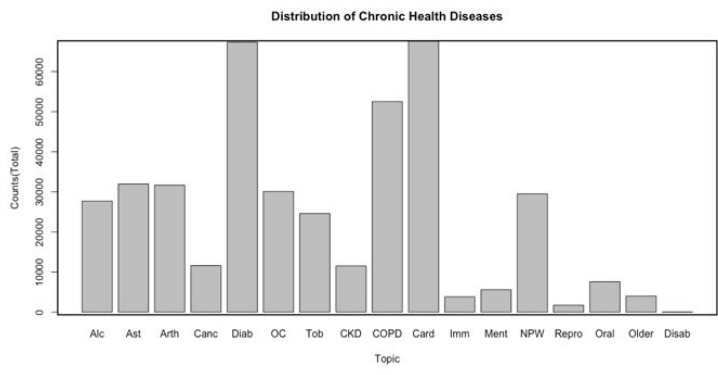


Figure 6. Distribution of Chronic Health Diseases

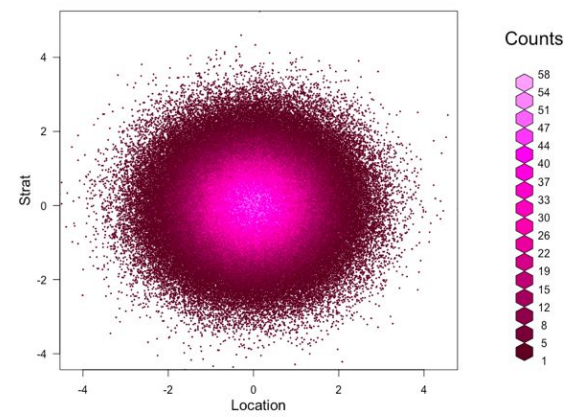


Figure 7. Hexagonal Binning showing Overlap between Location and Stratification.

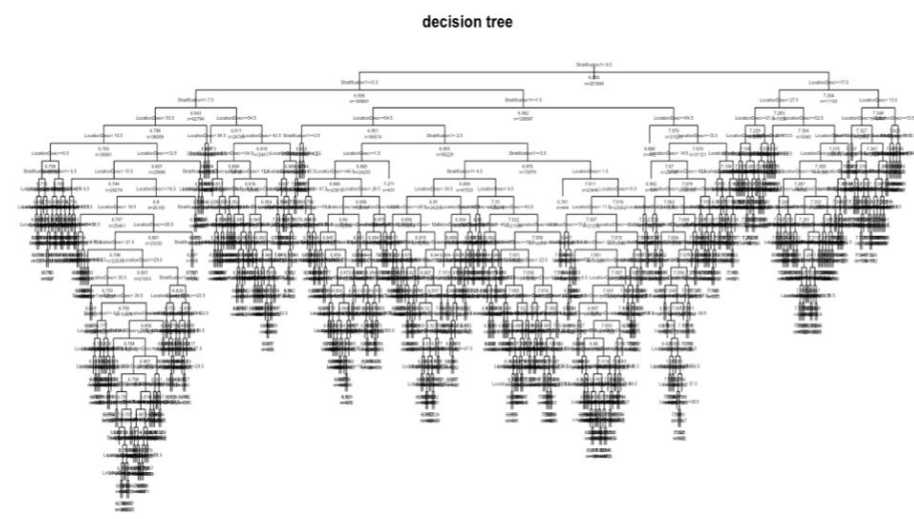


Figure 8. Decision Tree


```
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 9. 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 0 186 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 10. Prediction and Confusion Matrix

3 b. COVID-19 Dataset

Based on my distribution charts, there is a very equal distribution of Age, Ethnicity and Location in the data, which implies there might not have been random sampling (Figure 11, 12, 13). However, based on my scatterplot comparing death counts based on Age and Ethnicity (Figure 14) we can see that there is a much higher death count for minority ethnic groups when compared to Caucasians. In the density plot, looking at the overlap of Age and Ethnicity (Figure 15), there is not much overlap, and where the overlap is could be the minority ethnic groups and the older age groups. I created three different decision trees, the first one predicting probability of death based on age and ethnicity (Figure 16), the second one predicting death based on age (Figure 17) and the last on predicting death based on ethnicity (Figure 18). These decision trees are quite clear and provide the

expected probabilities. The random forest results (Figure 19) are also much cleaner and provide a better prediction of probability of death. Figure 20 shows an example of the prediction for death based on ethnicity. Since I suspected bias in data collection, I also wanted to see how much importance was given to each of my predictor variables, and the importance plot (Figure 21) shows that both variables were similar in treatment, which proves my theory of there being a bias.

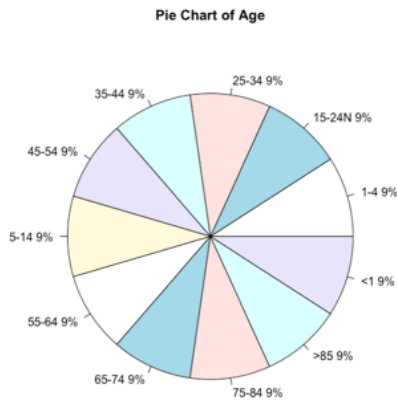


Figure 11. Age Distribution

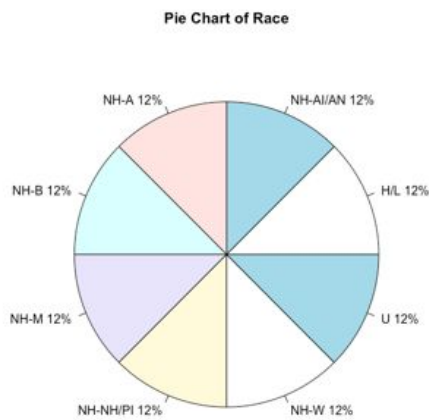


Figure 12. Ethnicity Distribution

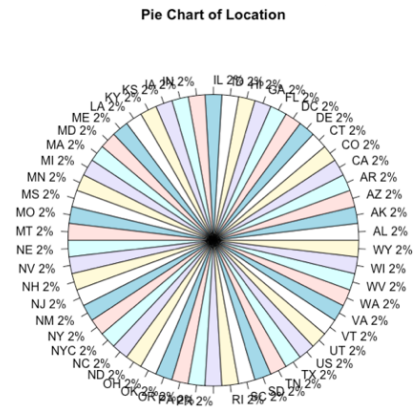


Figure 13. Location Distribution

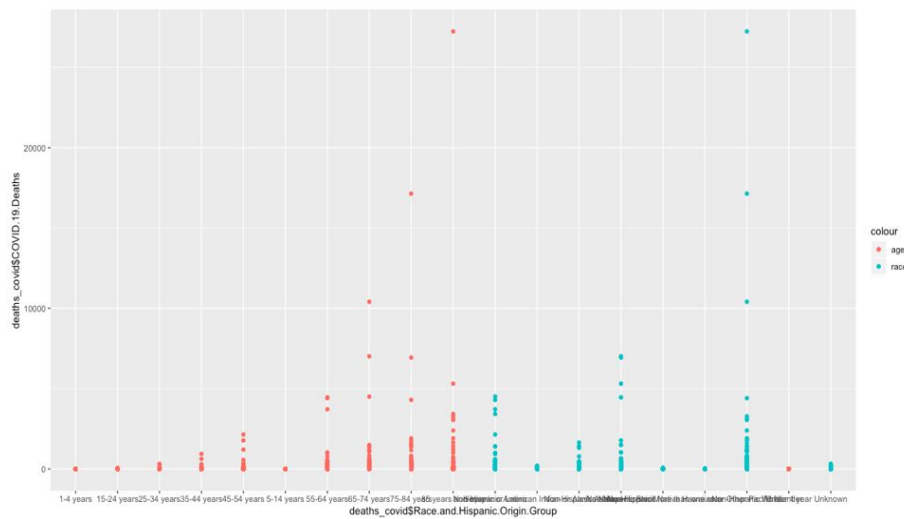


Figure 14. Scatterplot Comparing Age and Ethnicity Death Counts

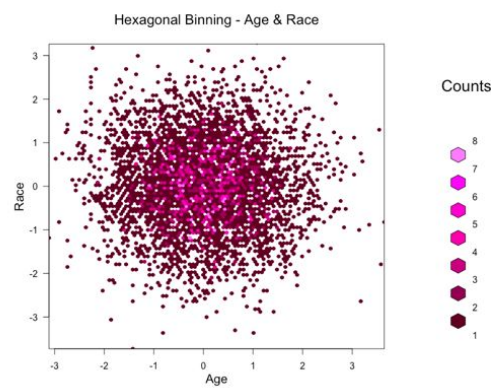
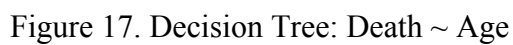


Figure 15. Density Plot of Age and Ethnicity



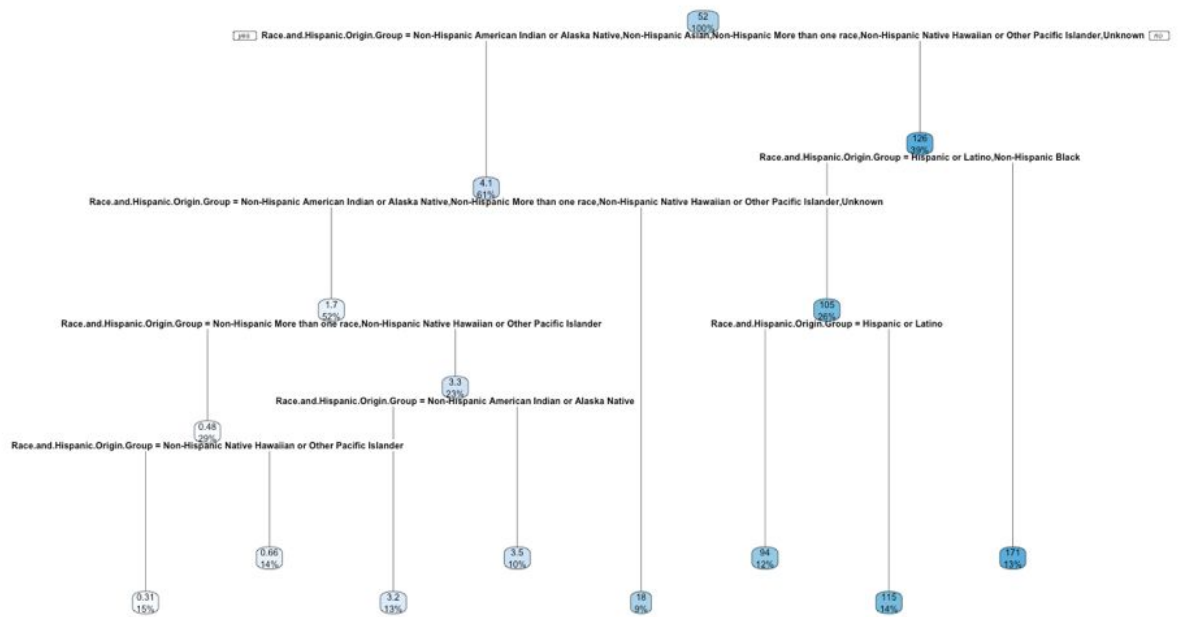


Figure 18. Decision Tree: Death ~ Ethnicity

<p>Call:</p> <pre>randomForest(formula = COVID.19.Deaths ~ Age.group + Race.and.Hispanic.Origin.Group, data = train)</pre> <p>Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 1</p> <p>Mean of squared residuals: 694985 % Var explained: 4.42</p>	
<p>Call:</p> <pre>randomForest(formula = COVID.19.Deaths ~ Age.group, data = train)</pre> <p>Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 1</p> <p>Mean of squared residuals: 715380.1 % Var explained: 1.61</p>	
<p>Call:</p> <pre>randomForest(formula = COVID.19.Deaths ~ Race.and.Hispanic.Origin.Group, data = train)</pre> <p>Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 1</p> <p>Mean of squared residuals: 711443.5 % Var explained: 2.15</p>	

Figure 19. Random Forest Results for each situation

```

> prediction_for_table <- predict(rf2,deaths_covid[,-4])
> table(observed=deaths_covid[,3],predicted=prediction_for_table)

```

observed	predicted	
Hispanic or Latino	0.0336632700420626	0.0337124805989236
Non-Hispanic American Indian or Alaska Native	0	0
Non-Hispanic Asian	0	0
Non-Hispanic Black	0	0
Non-Hispanic More than one race	0	0
Non-Hispanic Native Hawaiian or Other Pacific Islander	54	54
Non-Hispanic White	0	0
Total Deaths	0	0
Unknown	0	0

observed	predicted	
Hispanic or Latino	0.0383692496864414	0.0384838475922691
Non-Hispanic American Indian or Alaska Native	0	0
Non-Hispanic Asian	0	0
Non-Hispanic Black	0	0
Non-Hispanic More than one race	54	54
Non-Hispanic Native Hawaiian or Other Pacific Islander	0	0
Non-Hispanic White	0	0
Total Deaths	0	0
Unknown	0	0

observed	predicted	
Hispanic or Latino	0.0401038302645866	0.0506806789046423
Non-Hispanic American Indian or Alaska Native	0	0

Figure 20. Sample Confusion Matrix of Prediction for Death based on Ethnicity

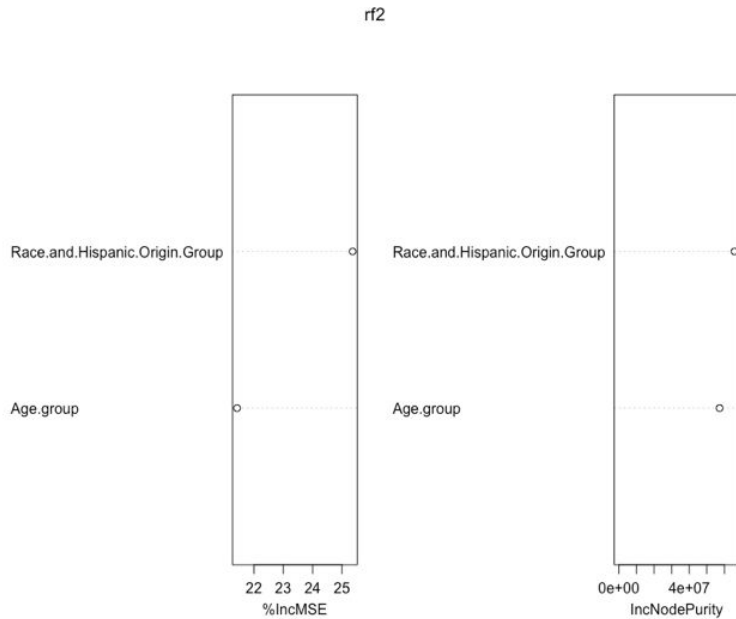


Figure 21. Importance Plot

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 applied this knowledge on my COVID dataset, and was able to 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. I also saw a similar trend in death counts due to COVID-19 where minority ethnic groups and older aged people had a much higher death count. (Figure 22, 23). I could further apply this understanding into looking at how other pandemics like Pneumonia and Influenza disproportionately affect minority ethnic groups.

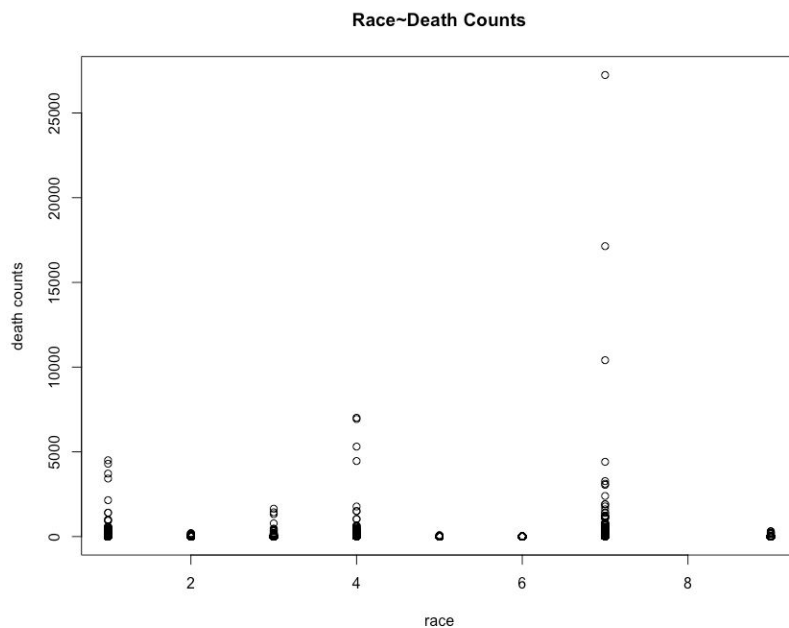


Figure 22. Death counts due to COVID-19 based on race

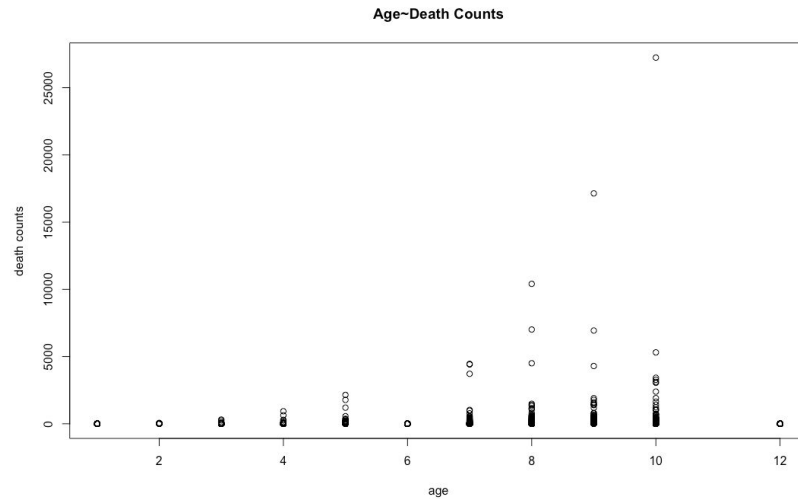


Figure 23. Death counts due to COVID-19 based on Age

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. Kirby, Tony. "Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities." *The Lancet. Respiratory medicine* vol. 8,6 (2020): 547-548. doi:10.1016/S2213-2600(20)30228-9
4. Davies, Nicholas G., et al. "Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics." *Nature News*, Nature Publishing Group, 16 June 2020, www.nature.com/articles/s41591-020-0962-9.

5. Centers for Disease Control and Prevention. "Chronic Disease Indicators." *Kaggle*, 17 Aug. 2017, www.kaggle.com/cdc/chronic-disease.
6. "Provisional COVID-19 Death Counts by Sex, Age, and State." *Data.gov*, Publisher Centers for Disease Control and Prevention, 25 June 2020, catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state.
7. Deng, Houtao. "Why Random Forests Outperform Decision Trees." *Medium*, Towards Data Science, 12 Dec. 2018, towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5.