



STEVENS
INSTITUTE *of* TECHNOLOGY
THE INNOVATION UNIVERSITY®

Predictive model for the US county-level diffusion of COVID-19

Artificial Intelligence & Machine Learning Approach

Vinay Rajesh Gor

Introduction

Coronavirus disease 2019 (COVID-19) is caused by a new virus identified first in Wuhan China in December 2019 started spreading around the world in a few days. Till October around 46 million people were confirmed with the virus causing a death of 1.2 million globally. The United States of America had the most number of cases accounting for 9.25 million with 231,000 deaths.

Project Goals

The purpose of this project is to create a predictive model for the US county-level diffusion of COVID-19. This will help to understand how various factors such as Total Population, Age, Housing Units in each county, Age and race wise distribution affect COVID-19 at county level.

Business Understanding

The objective of this study is to predict COVID-19 cases and deaths by county as a Data Scientist to determine the various factors which can directly or indirectly affect the people.

This study can help us to understand,

1. What is the trend of the COVID-19 county wise
2. Top counties with most number of cases
3. Top counties with most number of deaths
4. What factors are responsible for number of cases (or deaths)
5. Decision tree for the target variable with a threshold value related to COVID-19

The results obtained from this study can be analyzed by the Health Care Administration, Government, Politicians, Healthcare Analyst, Policy Makers to understand various aspects related to COVID-19.

Different data obtained will be in *.csv (comma separated values)* which will be merged into a single file *.csv (comma separated values)*. All the data exploration, visualization, modelling, evaluation will be performed in *Python* using different libraries such as *Pandas, Numpy, Seaborn, Matplotlib, Sklearn etc.*

Data Understanding

The data is a set of combinations of different .csv (comma separated values) from different sources which included year wise GDP, year wise Housing Units, year wise Census and Total number of cases & death by each county.

We are going to work on the most recent data hence all the data related to GDP, Housing units and Census from previous year has been excluded. Only recent data of recent years has been taken.

COVID-19 cases - 644679 rows x 4 columns

Census - 716377 rows x 80 rows

Housing Units - 3153 rows x 13 rows

GPD - 3223 rows x 10 columns

We see that as our data is not of the same dimensions we will first need to merge all files in a proper dimension which will cover all factors for all counties.

Data Preparation

From the data understanding we know that our data is not of the same dimensions hence we will need to merge all the files in a proper same dimension so that we can fit all data for all counties to be analyzed in Python.

The following things were done for data preparation:-

1. For COVID-19 cases only the recent data from October 18 2020 is taken which includes the total number of cases, deaths for each county of all states in the US. This is done using datetime function in python (*date == '2020-10-18'*)
2. For Census, Housing Units and GDP only the data from the most recent year is taken and the rest others are filtered or either deleted
3. All files are merged in python using function *.merge* based on combination of and/or *County, State & Fips code*
4. *All unknown values, non matching values, NAN values were dropped off*

```
#merging GDP & COVID cases file
df_covid_gdp = pd.merge(oct18,df_gdp,on=["county", "state"])
print(df_covid_gdp)
```

```
#merging Census, Housing, GDP, and COVID cases file
df_covid_gdp_housing_census = pd.merge(df_covid_gdp_housing,df_census,on=["county","fips"])
print(df_covid_gdp_housing_census)
```

Fig. Files merged based on combination of county, state, fips

Final Data Frame

Finally, a data frame of 3069 rows x 15 columns was generated which included the number of cases & deaths for each county as of 18th October 2020 with GDP, Housing, State, Fips code, Total population, and age groups(divided as Child & Adolescence, Young Adult, Mature Adult, Senior citizens).

	date	county	state_x	fips	cases	deaths	GDP
0	2020-10-18	Autauga	Alabama	1001.0	1989	28	1483414.0
1	2020-10-18	Baldwin	Alabama	1003.0	6369	67	5774289.0
2	2020-10-18	Barbour	Alabama	1005.0	981	9	787425.0
3	2020-10-18	Bibb	Alabama	1007.0	785	13	364197.0
4	2020-10-18	Blount	Alabama	1009.0	1827	23	849114.0
...
3065	2020-10-18	Sweetwater	Wyoming	56037.0	415	2	3836603.0
3066	2020-10-18	Teton	Wyoming	56039.0	700	1	2166420.0
3067	2020-10-18	Uinta	Wyoming	56041.0	401	2	906587.0
3068	2020-10-18	Washakie	Wyoming	56043.0	133	7	358104.0
3069	2020-10-18	Weston	Wyoming	56045.0	93	0	315885.0

	Housing Units	state_y	Total Population	Chil & Adolescence \
0	23896	Alabama	55869	14252
1	119412	Alabama	223234	52268
2	12080	Alabama	24686	5595
3	9261	Alabama	22394	4992
4	24517	Alabama	57826	14522
...
3065	19909	Wyoming	42343	12049
3066	14186	Wyoming	23464	4586
3067	9108	Wyoming	20226	6215
3068	3868	Wyoming	7805	1960
3069	3568	Wyoming	6927	1529

	Young Adult	Matured Adult	Senior Citizens
0	14171	15152	12294
1	49706	58282	62978
2	6486	6204	6401
3	6200	6079	5123
4	13666	15164	14474
...
3065	11580	10362	8352
3066	7212	6377	5289
3067	4770	4774	4467
3068	1613	1910	2322
3069	1607	1735	2056

[3070 rows x 15 columns]

Fig. Final dataframe

Creating Target Variable (Outcome)

In order to model our study a target variable needs to be decided which will help us to analyze the outcome and factors associated with it.

Cases per 100 population which implies the number of cases happened in a county per 100 people of that county or simply the percentage of people with COVID-19 virus in a county.

$$\text{Cases per 100 population} = \frac{\text{Number of COVID-19 cases}}{\text{Total population}} \times 100$$

Outcome will be the threshold value of our data.

Any county which has 5 cases per 100 population will be considered as a red zone having more than usual number of cases and will be converted to boolean value 1 or otherwise 0 (1 - if cases per 100 population > 5 (county not safe) or 0 (county safe)).

Exploratory Data Analysis

Descriptive statistics

First initial statistical data analysis is done to analyze the numeric characteristics of our data. This is achieved using *describe()* function.

	cases	deaths	GDP	Housing Units	Total Population	Chil & Adolescence	Young Adult	Matured Adult	Senior Citizens	cases/100 population	deaths/100 population
count	3070.000000	3070.000000	3.070000e+03	3.070000e+03	3.070000e+03	3.070000e+03	3.070000e+03	3.070000e+03	3.070000e+03	3070.000000	3070.000000
mean	2498.311401	62.062215	5.562040e+06	4.331924e+04	1.018617e+05	2.539727e+04	2.753034e+04	2.567617e+04	2.325792e+04	2.385831	0.049653
std	9735.045893	261.803897	2.325163e+07	1.276090e+05	3.288387e+05	8.239253e+04	9.648309e+04	8.429344e+04	6.734244e+04	1.588268	0.056022
min	1.000000	0.000000	1.835600e+04	1.860000e+02	2.720000e+02	7.400000e+01	6.000000e+01	6.400000e+01	7.400000e+01	0.057208	0.000000
25%	182.500000	2.000000	3.669698e+05	5.550000e+03	1.080300e+04	2.604250e+03	2.465500e+03	2.633500e+03	2.987500e+03	1.287482	0.012008
50%	540.000000	9.000000	9.497080e+05	1.258750e+04	2.563000e+04	6.273000e+03	6.055500e+03	6.423500e+03	6.785000e+03	2.107328	0.032369
75%	1543.500000	32.000000	2.700602e+06	3.150200e+04	6.704850e+04	1.664125e+04	1.692400e+04	1.690450e+04	1.738200e+04	3.160675	0.067139
max	288451.000000	6876.000000	7.108933e+08	3.579329e+06	1.003911e+07	2.399105e+06	3.035259e+06	2.619522e+06	1.985221e+06	17.037776	0.520279

Fig. Descriptive statistics

Statistical analysis of all variables is analyzed.

For a county the number of cases ranges from 1 to 2,88,451 whereas deaths countywise ranges from 0 to 6,876.

Cases per 100 population ranges from 0.05 to 17 cases per 100 population with an average of 2.

The mean GDP is 5.5 million dollars and the mean housing units is 40,000.

The average death rate across all counties is 4 per 10000 population.

Population wise maximum there are maximum Young adults and the least Senior citizens.

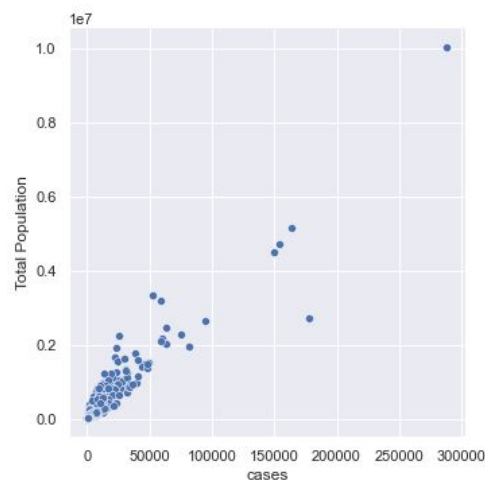


Fig. Linear relation between Total population & Cases

Correlation Matrix

Correlation is the study of numeric relation between variables. This forms an important aspect to determine how one factor is related to the other. This is achieved by corr() function.

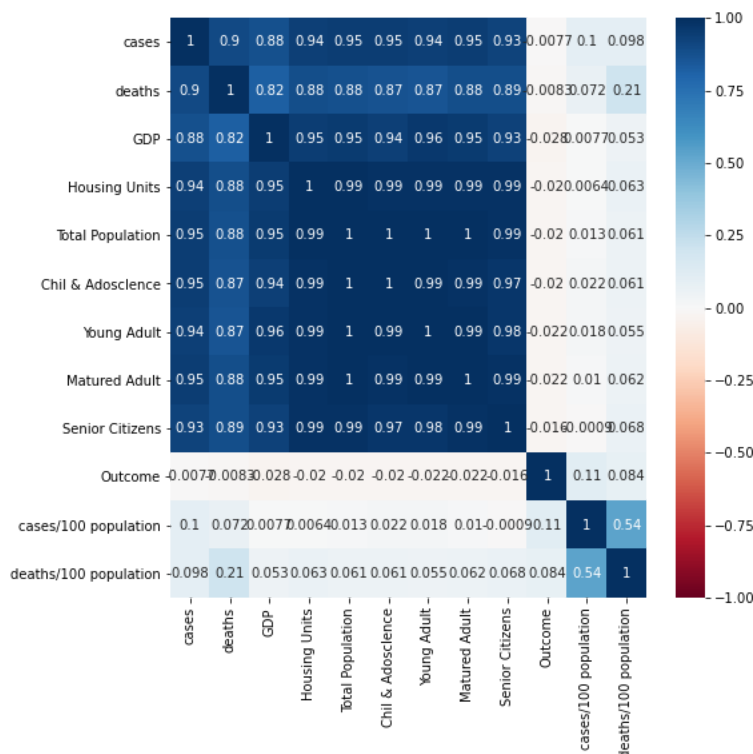


Fig. Correlation Matrix

It can be clearly visualized that the matrix consists of two colours red and blue. Red indicates the variables are negatively correlated to each other whereas Blue indicates the variables are positively correlated to each other.

Correlation matrix shows that cases are mostly related to Total population which clearly states that the more population and the more activities the more will be the cases and this is clearly evident across states like New York, New Jersey, California, Texas which have more cases. Cases have the same effects mostly on all ages. Cases per 100 population has most effect on Child and Adolescent whereas deaths per 100 population have most effect on Senior citizens. This is justified because most of the people who died were senior citizens.

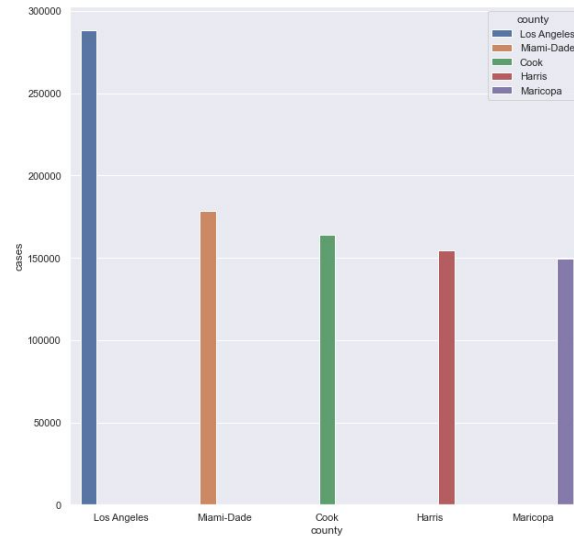


Fig. Maximum number of cases

The maximum number of cases as of October 18 2020 happened in counties Los Angeles with more than 250,000 cases whereas the next in list were Miami-Dade, Cook county, Harris, and Maricopa each counting for more than 150,000 cases.

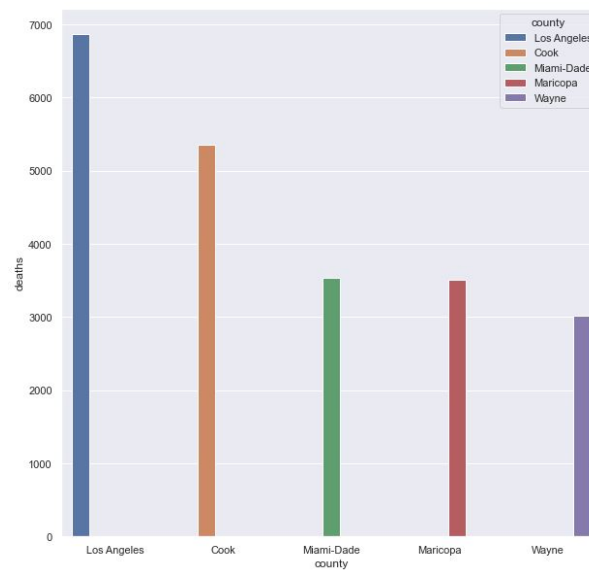


Fig. Maximum number of cases

The maximum deaths happened in Los Angeles and Cook as of October 18 2020 whereas next in the list were Miami-Dade, Maricopa and Wayne.

Modeling

Decision Tree

Decision tree is a machine learning predictive modelling technique. It is used to observe the decision factors and make a conclusion based on the target variable. The target variable/Outcome set here is cases per 100 population which is a binary value. Decision Tree was created using libraries *sklearn* and *graphviz*.

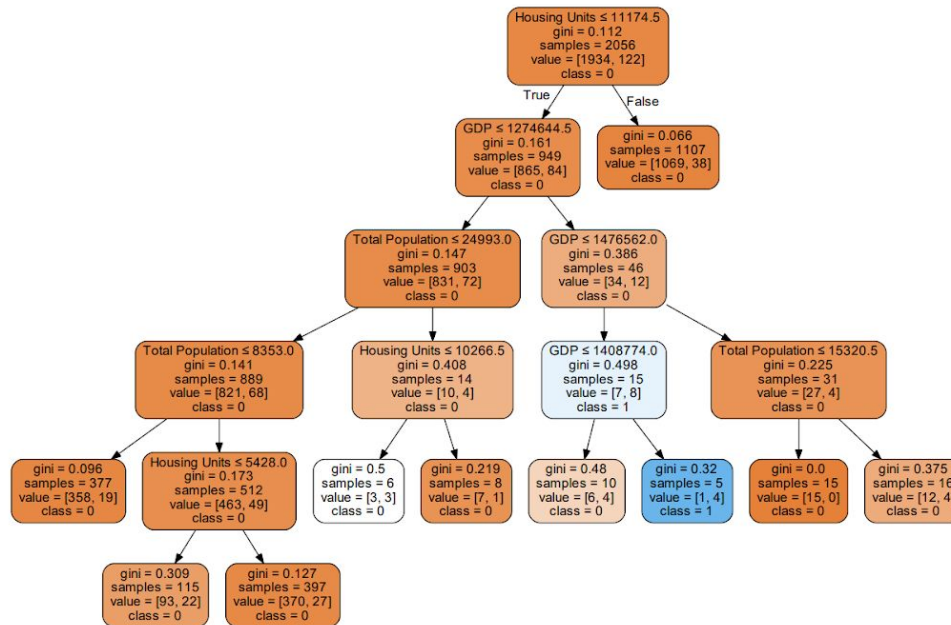


Fig. Decision tree

Causes of condition

From all the above figures we can analyze using the Decision tree that the most important factor is the number of Housing Units which is the most pure factor with a gini index of 0.112. This logically makes sense if the county is having more number of Housing units then less density of people living across the county whereas less number of houses reflect more density and hence more spreading of COVID-19.

The other two factors which are next important is GDP and Total population which indicates the more population the more number of people and more chances of spreading virus.

Talking about Age when analyzed Senior citizens had the most pure factor.

Random Forest

Random is a predictive machine learning technique which can be used for Classification, Regression, Observation using many decision trees and outputting the mode/mean/average of individual trees. Random forest corrects decision trees for overfitting and generally random forests outperform decision trees. In random forest the decision trees are split randomly and each time every other feature is considered for splitting which creates more diversification.

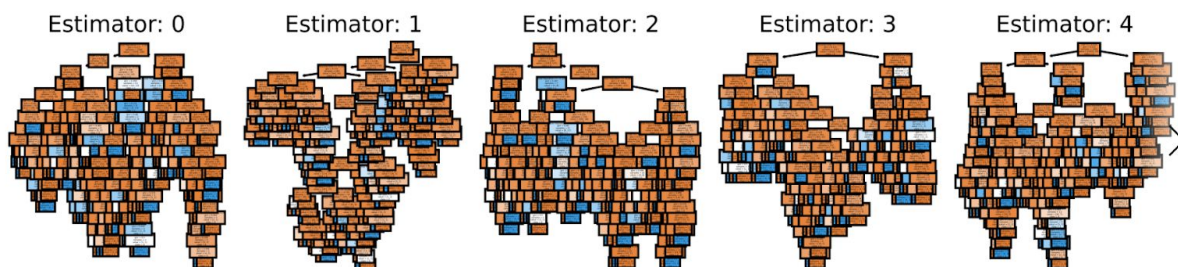


Fig. Different decision trees created by Random Forest

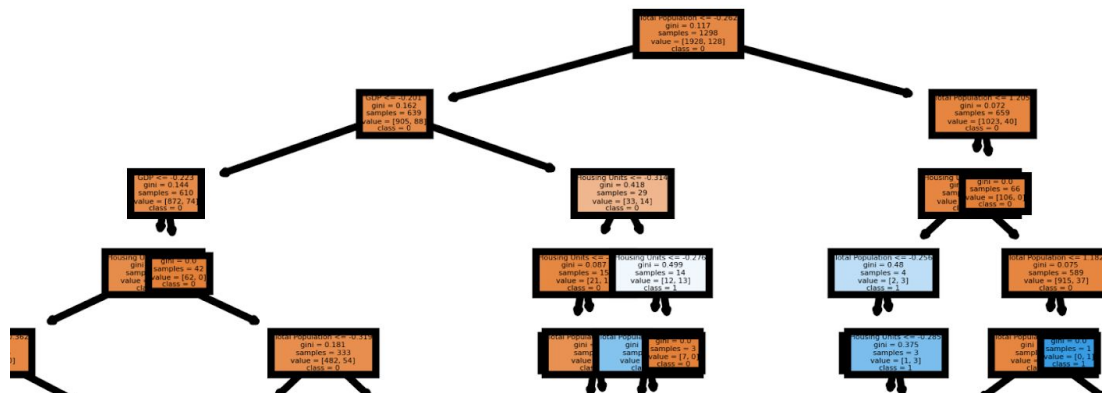
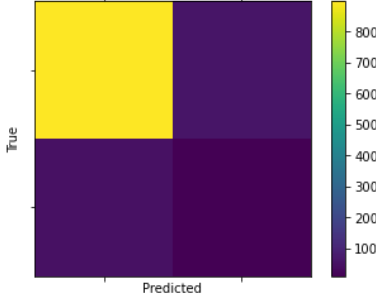
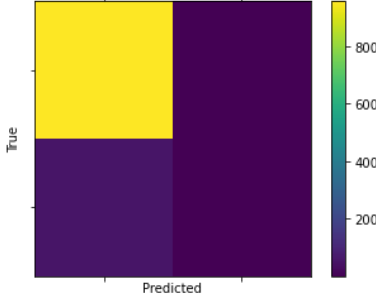


Fig. Random forest

Random forest created different decision trees splitting by each feature and based on that finally created the final decision tree which is the mean of all the decision trees which is Total population with a gini index of 0.117 which implies that more the population the more is the number of cases per 100 population. The next factors responsible are GDP and then Housing.

Evaluation

Factors	Decision Tree	Random Forest
Gini Index	0.112	0.117
Most Important Factor	Housing Units	Total Population
Accuracy	<pre>bc_tree.score(X_test, Y_test)</pre> 0.893491124260355	<pre>rfcmodel.score(X_test, Y_test)</pre> 0.9447731755424064
Confusion Matrix	<p>Confusion matrix of the Classifier</p> 	<p>Confusion matrix of the Classifier</p> 
Confusion Matrix	<pre>[[897 60] [48 9]]</pre> <p>True Negative - 897 True Positive - 9 False Negative - 60 False Positive - 48</p>	<pre>[[956 1] [55 2]]</pre> <p>True Negative - 956 True Positive - 2 False Negative - 1 False Positive - 55</p>

Decision Tree

The table shows that out of 1014 samples 69 were predicted the county to be in the red zone and 945 were predicted to be in the green zone. Out of which in actual reality 57 were actually observed to be red zone and 957 to be in green zone.

Accuracy = (True Positive + True Negative) / Total = 89.34%

Random Forest

The table shows that out of 1014 samples 3 were predicted the county to be in the red zone and 1011 were predicted to be in the green zone. Out of which in actual reality 57 were actually observed to be red zone and 957 to be in green zone.

Accuracy = (True Positive + True Negative) / Total = 94.47%

The comparison above shows that both models are having good accuracy but Random Forest outperforms Decision trees. This obviously depends on the other factors.

Deployment

The results created can be used by the organizations such as to improve the situation created by Demonetization by the Health Care Administration, Government, Politicians, Healthcare Analyst, Policy Makers to understand aspects related to COVID-19 such as Total Population, Housing, GDP, age groups such as Senior citizens. Some of the factors or all factors can be taken into consideration and worked upon as COVID-19 is still not over yet with no vaccine around.

Conclusions

To conclude this study helped us to predict the US county-level diffusion of COVID-19 where we distributed, analyzed the trend of COVID-19 county wise cases and deaths for the most recent dates. We determined the maximum number of cases and deaths for counties in the US. Using correlation, models & algorithms such as Decision Tree and Random Forest we determined factors or hierarchy of factors responsible for COVID-19 using a target variable case per 100 population.

Random forest performance was better than decision trees as a predictive classification model.

Improvisation & Future Works

Other factors such as number of people cured, Healthcare facilities, Immunity, Income, Hospitalization, area of county etc could make our predictive classification model more efficient and reliable. For future works I would like to explore more on the above mentioned factors.

.