CSE3505 - FOUNDATIONS OF DATA ANALYTICS

J Component Report

A project report titled

Corona virus and wealth

By

Reg.no: 17MIS1022 Name: K. Karthik

Reg.no: 17MIS1057 Name: G. Adithya Sai

M.Tech (Software engineering)

Submitted to

Dr. R. Rajalakshmi

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

November 2020

DECLARATION BY THE CANDIDATE

I hereby declare that the report titled "CoronaVirus and Wealth" submitted by me to VIT Chennai is a record of bona-fide work undertaken by me under the supervision of **Dr. R. Rajalakshmi, Associate Professor, SCOPE, Vellore Institute of Technology, Chennai.**

Signature of the Candidate

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Jagadeesh Kannan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report entitled "**CoronaVirus and Wealth"** is a bona-fide work of **K. Karthik (17MIS1022), G. Adithya Sai (17MIS1057)** carried out the "J"-Project work under my supervision and guidance for CSE3505 - Foundations of data analytics.

Dr. R. Rajalakshmi

SCOPE

TABLE OF CONTENTS

Abstract:

The government agency responsible for the whole-of-government approach has collected data by state and county regarding population wealth and COVID cases and deaths for two points in time during the pandemic. The agency is interested to know if population wealth is an indicator of per capita COVID cases and deaths, death rate from confirmed cases, rate of change of these factors, and spread to adjacent principalities.At both the state and county levels, local governments could better prepare, warn citizens, and tighten recommended and/or required preventative measures One hypothesis, among many, is that population demographics related to wealth might provide insight into COVID-19 spread. Machine Learning algorithms such as Random Forest, SVM (Linear Kernel), KNN.But there are some inconsistencies that are to be taken care while classifying. One of them includes several features present in the dataset. We have used PCA analysis to under the level of feature extraction ,mutate columns and aggregate to make more reliable features.This analysis is focused on one of the questions: is population wealth an indicator of death rate from confirmed COVID cases.

Introduction:

Federal, State and County government officials are developing a whole-of-government approach to COVID-19 infection prevention and management. Key to such an approach is an understanding of where COVID-19 is most likely to spread, most likely to spread the fastest, and cause the most death. At the federal level, having this understanding would allow better allocation of nationally controlled resources such as testing kits, emergency reserve and newly manufactured ventilators, and military medical augmentation. At the state level, medical assets such as people (doctors, nurses and respirator technical staff), medical equipment, beds, supplies and medicine could be allocated proactively to predicted hotspots across counties and even across state lines where counties with similar demographics border one another. At both the state and county levels, local governments could better prepare, warn citizens, and tighten recommended and/or required preventative measures One hypothesis, among many, is that population demographics related to wealth might provide insight into COVID-19 spread. The government agency responsible for the whole-of-government approach has collected data by state and county regarding population wealth and COVID cases and deaths for two points in time during the pandemic. The agency is interested

to know if population wealth is an indicator of per capita COVID cases and deaths, death rate from confirmed cases, rate of change of these factors, and spread to adjacent principalities. This analysis is focused on one of the questions: is population wealth an indicator of death rate from confirmed COVID cases.

Requirements:

The following libraries are requirements: Tidyverse, forcats, ggplot2, dplyr, stargazer, caret, modelr, Hmisc, DataExplorer, usmap, skimr, devtools, cluster, visdat, tidyr, readr, arules, mltools, arulesViz, factoextra, stats, sqldf, MASS, klaR, e1071, rpart, rattle, rpart.plot, RColorBrewer, class, gmodels, FactoMineR.

Dataset Description:

Dataset aims to further a county by county analysis of potential risk factors that could heighten Covid 19 transmission rates or deaths. The data has now been split between general population and over
60 estimates and converted to counts for ease of use.
Dataset:(133 Features)
https://data.census.gov/cedsci/table?q=United%20States&tid=ACSDP1Y2018.DP05&hidePreview=true
https://www.census.gov/programs-surveys/acs/guidance/estimates.html
It includes a subset of county by county ACS estimates of:
The data includes information on:
1) County level indicators for over 60 populations including population density, race, poverty level,
housing size, sources of income, employment status, whether living alone, language barriers, immigration
status, and disability status.

2) County level indicators for the general population including race, poverty level, housing size, sources of
income, employment status, whether living alone, language barriers, immigration status, and disability
status, modes of transportation stats, and industry stats.

Problem definition:

The problem for the government is to find the coronavirus cases in the counties and where the virus most likely to spread fastly and to find the death rates. They should also allocate the emergency calls, medical equipment. They want to know that there is any relation between population and covid cases and deaths and which county is most affected by the virus.

Proposed System/ Modules description:

The data includes information on: 1) County level indicators for over 60 populations including population density, race, poverty level, housing size, sources of income, employment status, whether living alone, language barriers, immigration status, and disability status. 2) County level indicators for the general population including race, poverty level, housing size, sources of income, employment status, whether living alone, language barriers, immigration status, and disability status, modes of transportation stats, and industry stats.

Analysis and Models:

No. Of rows: 1476

No. Of columns(features): 133

County level indicators for over 60 populations including population density, race, poverty level, housing size, sources of income, employment status, whether living alone, language barriers, immigration status, and disability status.

Data Cleaning:

Data Cleaning is the process of transforming raw data into consistent data that can be analyzed. The aim is to improve the content of questions based on the data available. To improve the quality of the data for applying the techniques to analyse the data and to preprocess, we removed all the NA values from the dataset.

Aggregation of Rows:

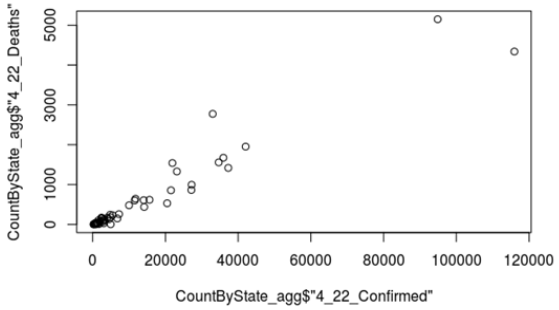
Aggregate() Function in R Splits the data into subsets, computes summary statistics for each subsets and returns the result in a group by form.Aggregate() function is useful in performing all the aggregate operations like sum,count,mean, minimum and Maximum. Aggregated the rows using SUM function by states

To know the maximum confirmed cases in the states we sorted the data in descending order. Sorted the data by using the "ORDER" function. Prepended the sorting variable by a minus sign to indicate descending order.

To know which state is affected by the virus, we have used USmap for visualising the data that which state is most affected by using the colors red and white. Red defines high cases and white defines low cases in the state.

To compare the stats of confirmed and death cases over day by day we have used a bar plot.

To know how many cases are confirmed and how many people who are dead.



Box plots are very effective and easy to read, as they can summarize data from multiple sources and display the results in a single **graph**. Displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

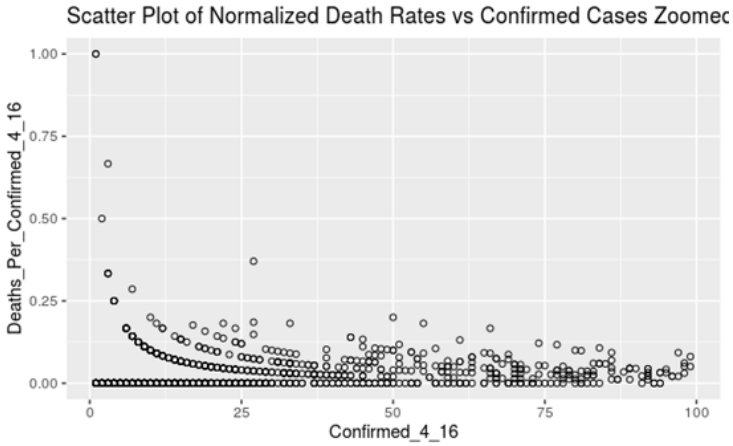
To know the minimum, maximum, first quartile, median and third quartile box plot can be used for this. Box plot for confirmed cases on 4_22.

For pre-processing the data, we have dropped the NA values.

Pre-processing:

We have mutated the 24 new columns to know the percentage between the features like how many percent people are black, white, under age, employed, graduates are there in the population who can get affected by coronavirus and we also saw the summarization of newly mutated 24 columns.

Plotted the graph between confirmed and deaths per confirmed where the confirmed cases are less than 100



We have calculated the Z Score of deaths per confirmed for knowing how many standard deviations away from the mean. Which the Z score is greater than 3 we have considered as an

outlier. And we want to know which county as death per confirmed cases greater than 0.25. The output gives the row numbers of the dataset.

```{r}
BigRatesRow <- which(New_Data$Deaths_Per_Confirmed_4_16>.25)
BigRatesRow

# Observation: There are 9/ 1476 data points with Death Rate > .25
```

```
[1]  122  278  408  560  593  728  876  917 1422
```

Machine Learning Models:

KNN with cross-fold Validation:

In this model building we have used KNN approach for classification of bins on the curated dataset with 19 significant features. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. KNN is coupled with repeated searchCV approach to find the optimal value of K and Kappa for the algorithm. This Algorithm has shown accuracy of ~49% and kappa score of 0.2. The search CV has chosen the K value as 43.
Observation: 25 values of K were attempted. K=43 was best with an training accuracy of 50% and kappa .18 The testing accuracy is 49% This model is putting most of the predictions in the first bin and a few in the 3rd and 4th and none elsewhere

SVM - Linear:

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.
The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified

correctly. Conversely, a very small value of C will cause the optimizer to look for a larger margin separating hyperplane, even if that hyperplane misclassified more points.
Observation: With a training accuracy of ~52%. The testing accuracy is 49%. Various randomised searchCV approach is designed to get the optimal value of C.

grid <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5))

Accuracy was used to select the optimal model using the largest value. The final value used for the model was C = 5.

PCA and Random Forest:

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

To perform principal component analysis on the dataset we need to convert the dataframe to the Matrix and on applying the PCA , we are able to deduce eigenvalues and vectors.This eigenvalues can be processed and plotted on the graph to identify the features that weight more under dimensionality apart from correlation analysis. RF is fed with the normalised death dataframe obtained from PCA analysis and then trained with the model of RF.

Observation: With a training accuracy of 47% that is performed on the cross validation of 10 fold, repeated 3 times.

The PCA analysis was able to reduce the dimension from 19 to 5 dimensions. This is fed with the features of DF (Death on 4_16) and then predicted.

Results and Discussion:

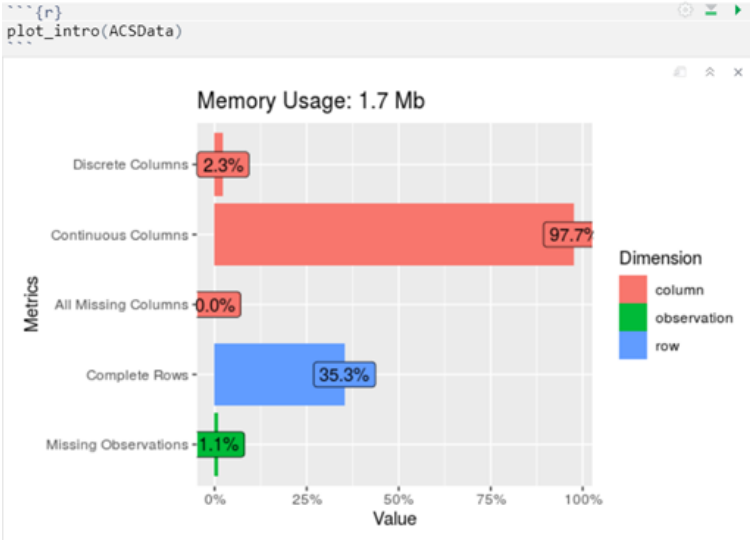


Figure : Basic information of input data

	state <fctr>	4_22_Confirmed <dbl>	4_22_Deaths <dbl>	4_16_Confirmed <dbl>
1	ALABAMA	4874	179	3848
2	ALASKA	282	5	253
3	ARIZONA	5471	231	4235
4	ARKANSAS	1361	33	1192
5	CALIFORNIA	37290	1418	27626
6	COLORADO	10001	479	7612

Figure : Aggregated data

	state <fctr>	Confirmed_Rise_4_16_to_20 <dbl>
31	NEW JERSEY	20600
33	NEW YORK	15445
22	MASSACHUSETTS	10545
5	CALIFORNIA	9664
14	ILLINOIS	9098
39	PENNSYLVANIA	7808
21	MARYLAND	6991
7	CONNECTICUT	6569
36	OHIO	5678
47	VIRGINIA	5206

1-10 of 51 rows Previous 1 2 3 4 5 6 Next

Figure : Sorted data in descending order

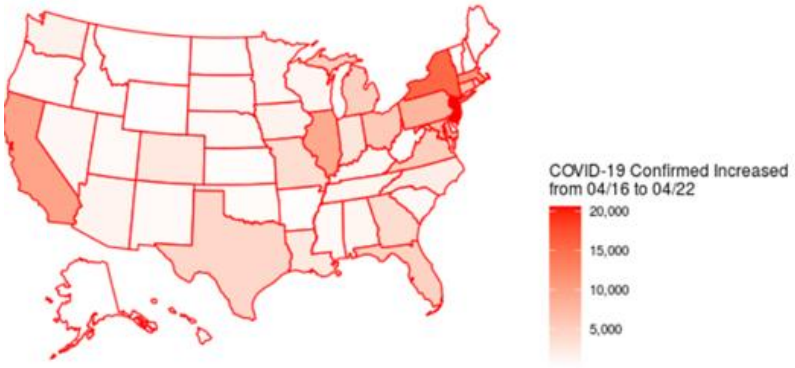


Figure : USmap for confirmed cases

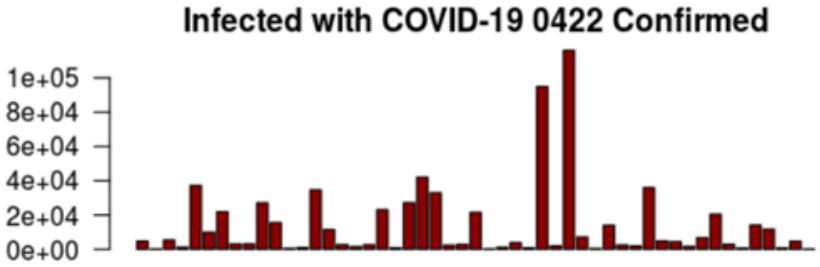


Figure : Active rise of Cases from 04-22

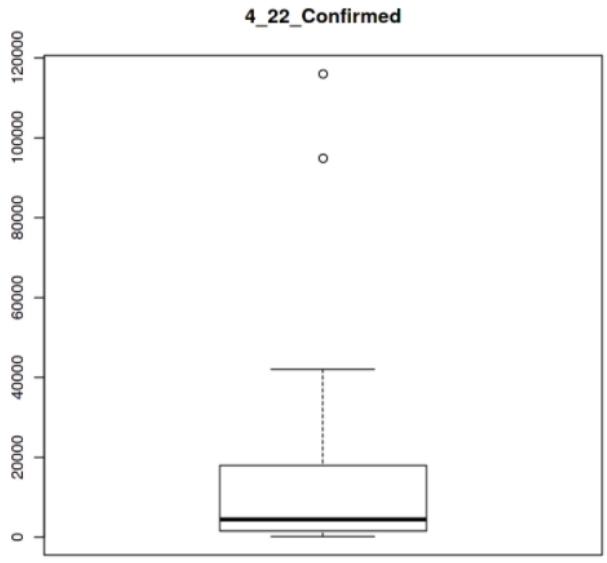


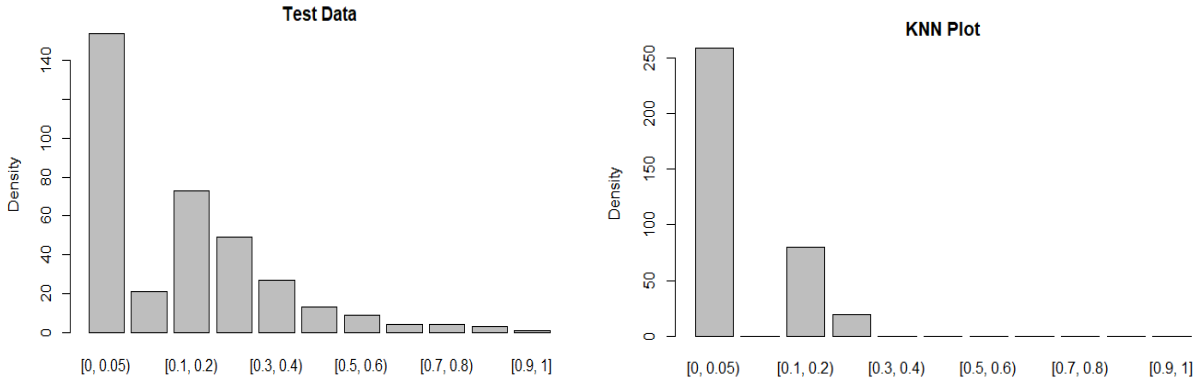
Figure : Box plot for confirmed cases on 4_22

```
       ID                  STATE             COUNTY      Pop_Density
1       :   1   TEXAS          :  96   Washington:  17   Min.   :    2.404
2       :   1   OHIO           :  75   Franklin  :  16   1st Qu.:   63.419
3       :   1   NORTH CAROLINA:  74   Jefferson :  16   Median :  119.569
4       :   1   VIRGINIA       :  62   Jackson   :  12   Mean   :  380.267
5       :   1   GEORGIA        :  60   Lincoln   :  12   3rd Qu.:  310.787
6       :   1   MICHIGAN       :  60   Montgomery:  12   Max.   :18565.536
(Other):1468   (Other)        :1047   (Other)   :1389
Confirmed_4_22     Deaths_4_22      Confirmed_4_16     Deaths_4_16
Min.   :    1.0   Min.   :   0.00   Min.   :    1.0   Min.   :   0.00
1st Qu.:   18.0   1st Qu.:   0.00   1st Qu.:   14.0   1st Qu.:   0.00
Median :   53.0   Median :   2.00   Median :   41.0   Median :   1.00
Mean   :  470.9   Mean   :  21.08   Mean   :  370.3   Mean   :  14.39
3rd Qu.:  195.0   3rd Qu.:   7.00   3rd Qu.:  146.0   3rd Qu.:   5.00
Max.   :31555.0   Max.   :1431.00   Max.   :27772.0   Max.   :1109.00

ET_Total_Population EM_Total_Pop_Median_Age Confirmed_Per_Capita_4_22
Min.   :   18699    Min.   :24.60           Min.   :1.578e-05
1st Qu.:   43016    1st Qu.:37.10           1st Qu.:3.263e-04
Median :   75242    Median :40.10           Median :6.378e-04
Mean   :  204250    Mean   :40.17           Mean   :1.327e-03
3rd Qu.:  178056    3rd Qu.:43.10           3rd Qu.:1.311e-03
Max.   :10098052    Max.   :67.00           Max.   :3.270e-02

Deaths_Per_Capita_4_22 Deaths_Per_Confirmed_4_22 Confirmed_Per_Capita_4_16
Min.   :0.000e+00      Min.   :0.00000           Min.   :1.578e-05
1st Qu.:0.000e+00      1st Qu.:0.00000           1st Qu.:2.538e-04
Median :2.057e-05      Median :0.02737           Median :4.963e-04
Mean   :5.629e-05      Mean   :0.04008           Mean   :1.001e-03
3rd Qu.:5.611e-05      3rd Qu.:0.05620           3rd Qu.:9.851e-04
Max.   :1.312e-03      Max.   :1.00000           Max.   :2.704e-02
```

```
Deaths_Per_Capita_4_16 Deaths_Per_Confirmed_4_16
Min.   :0.000e+00      Min.   :0.00000
1st Qu.:0.000e+00      1st Qu.:0.00000
Median :1.345e-05      Median :0.02255
Mean   :3.843e-05      Mean   :0.03500
3rd Qu.:3.737e-05      3rd Qu.:0.04928
Max.   :1.082e-03      Max.   :1.00000

Two_Week_Confirm_Rate_Per_Capita Two_Week_Death_Rate_Per_Capita
Min.   :-4.155e-05               Min.   :-1.621e-04
1st Qu.: 3.945e-05               1st Qu.: 0.000e+00
Median : 1.225e-04               Median : 0.000e+00
Mean   : 3.259e-04               Mean   : 1.785e-05
3rd Qu.: 3.145e-04               3rd Qu.: 2.041e-05
Max.   : 2.848e-02               Max.   : 4.367e-04

Fraction_Female  Fraction_White   Fraction_Black    Fraction_Other
Min.   :0.3880   Min.   :0.1500   Min.   :0.00100   Min.   :0.00700
1st Qu.:0.5010   1st Qu.:0.7522   1st Qu.:0.01400   1st Qu.:0.03800
Median :0.5070   Median :0.8740   Median :0.04200   Median :0.05900
Mean   :0.5053   Mean   :0.8209   Mean   :0.09599   Mean   :0.08312
3rd Qu.:0.5130   3rd Qu.:0.9310   3rd Qu.:0.12000   3rd Qu.:0.09500
Max.   :0.5390   Max.   :0.9840   Max.   :0.72300   Max.   :0.84200

   Hispanic       Population_Over_Age_15 Fraction_Less_Than_HS
Min.   :0.00500   Min.   :  14430        Min.   :0.0200
1st Qu.:0.02800   1st Qu.:  29481        1st Qu.:0.0850
Median :0.05300   Median :  50996        Median :0.1120
Mean   :0.09587   Mean   : 138100        Mean   :0.1204
3rd Qu.:0.10775   3rd Qu.: 119942        3rd Qu.:0.1460
Max.   :0.99100   Max.   :6845489        Max.   :0.4850
```

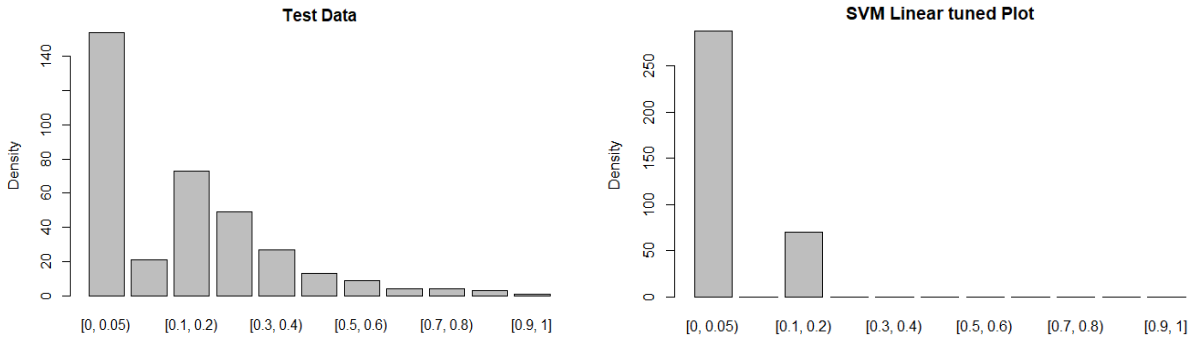
Figure : Summary of newly mutated data

KNN cross validation:



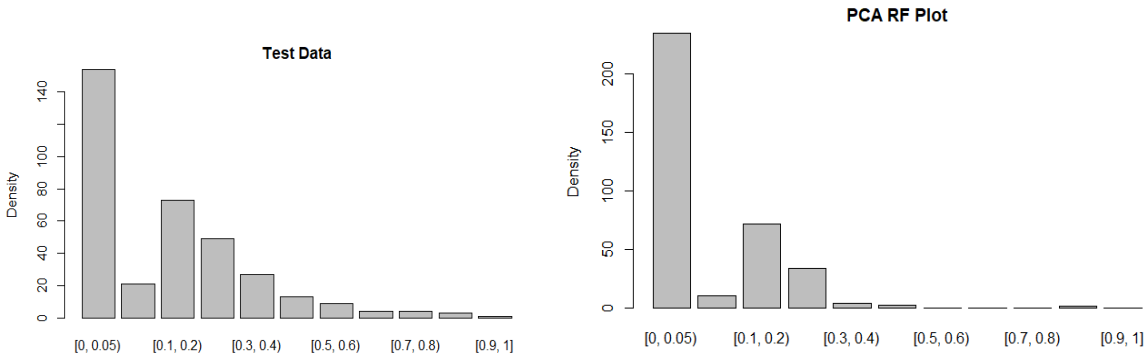
Left image contains the data distribution and the (right) model prediction.

SVM- Linear Kernel



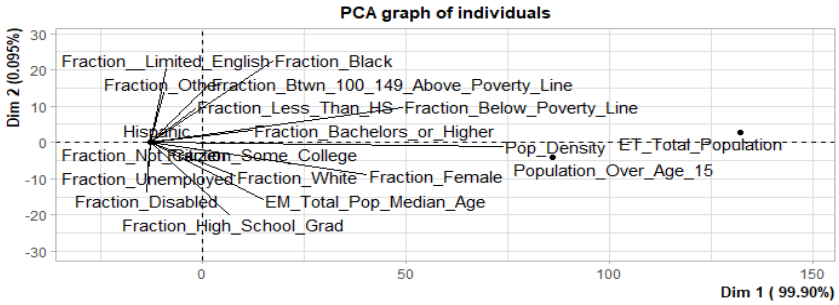
Left image contains the data distribution and the (right) model prediction.

PCA and RF implementation:



Left image contains the data distribution and the (right) model prediction.

PCA Curve:



Conclusion:

By the method of feature sampling we are able to achieve feature extraction and mutating new features from the dataset, which in turn reduces the dimensionality of data. We have explored new algorithms and compared them based on metrics like precision, recall, accuracy. PCA analysis for dimensional analysis has shown a significant reduction in features (in this case 5) and coupled with the Random forest Algorithm was able to achieve a good amount of accuracy on most features. This accuracy was then followed by SVM with the optimal values of C and then KNN (cv) approach with an optimal value of 41 were good in the prediction of the Death_cause of the county_ and future rise of the causes. As the US regions have multinational visits these features tend to help identify people with their resistance to race, gender, employment status of the individuals. Pandemics like coronavirus need to be carefully assessed, suppressed as soon as

possible, and spread of these pandemics need to be hindered. This level of feature analysis and prediction can affect early inhibition of the flow and save the economy, life of individuals with a statistical fool proof system.

Reference:

https://www.rdocumentation.org/
https://ggplot2.tidyverse.org/reference/ggplot.html
https://dplyr.tidyverse.org/
https://cran.r-project.org/web/packages/usmap/readme/README.html
https://cran.r-project.org/web/packages/skimr/vignettes/skimr.html
https://cran.r-project.org/web/packages/visdat/vignettes/using_visdat.html
https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix
https://www.rdocumentation.org/packages/caret/versions/4.47/topics/train
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/expand.grid