IST 707- DATA ANALYTICS

FALL 2020- FINAL PROJECT

EXPERIMENTING DATA ANALYTICS TECHNIQUES TO DETERMINE THE INFLUENCE OF GETTING IMPACTED WITH NOVEL CORONA VIRUS WITH PEOPLE WHO ARE HAVING CERTAIN EXISTING HEALTH ISSUES

SYRACUSE UNIVERSITY

PRASOONA KALLAGUNTA

CONTENTS

INTRODUCTION

Health is wealth as quoted by famous "Ralph Waldo Emerson" and this has been continuous saying of Mahatma Gandhi in multiple occasions. In the current 21st century, this generation is habituated to a fast forward life style and processed food rather than being believe on mother earth and nature and blessings of environment, climate conditions and so on. Hence we are killing our own immune system by paying lots of money.

Let me come to the current pandemic, this is something never ever thought through and we are busy in experiment how we can reach Mars, whether water existing is there on Jupiter and how many more suns are there in the space and how many more galaxies are existing, are aliens there? And we successfully failed to recognize the modern day "Bio War". However this gave us tremendous information(in terms of wide variety of data), various ways exploring alternatives, how to forecast trends, what are the loop holes in recording the data, do we have systems, processes in place to extract, explore and analyze millions and millions data in a sec. How best we can provide forecast to our people what are the trends?

Across the globe this pandemic impacted almost all kinds of living beings, and the worsen as day passes and finally the world came to a halt kind of mode, then our clinical research teams started their research with the collected data and coming up medication based on the age, existing conditions, demographical situations and so on.

PROJECT GOAL

My current goal is to figure a sample data that was collected during this pandemic and how COVID-19 is impacted with existing conditions to humans. Is there any relation among certain conditions? , by using the methods and techniques learnt as part of course (data mining algorithms, data analysis tools)

APPROACH

1. Collect the data either from Google, Kaggle
2. Perform Data cleanup
3. Identify the key pointers to analyze the data and extract the data (Partition, Split)
4. Interpret the data with best suitable models using R and or WEKA (with sample data)
5. Plot the graphs for better visualization

KEY METHODS

Below are the methods, models, techniques will be used to analyze, explore and interpret the data to accomplish the required results (result set)

- Predictive analytics
- Clustering
- Decision Trees(may not be able to fit with this kind of data)
- Naïve Bayes
- K-Nearest Neighbor
- Support Vector Machines

SAMPLE CODE

Below is the sample extract of code when tried to apply decision tree analysis after splitting the data to training and testing. However as predicted the results could not be visualized as anticipated due to large amount of column information.

R CODE:

```
######## FINAL PROJECT ######

# ## Clear objects from Memory
rm(list=ls())
### Clear Console:
cat("\014")
### Set Working Directory
setwd("E:\\PRASOONA\\MS_ADS\\IST707_DATA ANALYTICS\\HOMEWORK\\PROJECT")

## Install required packages
install.packages("caret")
library(caret)

## LOAD DATA
COVID19 <- read.csv("E://PRASOONA//MS_ADS//IST707_DATA
ANALYTICS//HOMEWORK//PROJECT/2020_country_daily_US_daily_symptoms_Subset.csv
")

## Split the data into training and testing
intrain <- createDataPartition(y=COVID19$date, p=0.7, list = FALSE)
training <- COVID19[intrain,]
testing <- COVID19[-intrain,]
```

```
dim(training)
str(training[, 1:10])

summary(training[, 10 :10])
head(training [, 1:10], n=5000)

# Build Decision Models
install.packages("rpart.plot")
library(rpart.plot)
install.packages("rpart")
library(rpart)
install.packages("e1071")
library(e1071)
install.packages("dplyr")
library(dplyr)
require(ggplot2)
install.packages("tree")
library(tree)

tree1 <- rpart(ï..sub_region_1 ~. - date, data = training, method = 'class', control =
rpart.control(cp=0))
summary(tree1)
rsq.rpart(tree1)
```

BELOW ARE THE PROBABILITY CAPTURED FOR ALL THE “48” STATES
Since the list is longer to compare for the 360 symptoms for all states .

I’ve considered only the key impacted states such as “New York, New Jersey, California.

```
code number 375142: 25 observations,    complexity param=0.0002326303
  predicted class=Maryland             expected loss=0.76  P(node) =0.001897389
    class counts:     0    0    0    0    0    0    0    0    0    1    2    0    1    1    0    0    1    0
0    0    6    0    0    0    0    1    0    0    0    0    1    0    0    3    1    0    0    0    2    0    0
0    0    1    0    0    4    0    0    0    0
   probabilities: 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.040 0.080 0.000 0.040
0.040 0.000 0.000 0.040 0.000 0.000 0.000 0.240 0.000 0.000 0.000 0.000 0.040 0.000 0.000
0.000 0.000 0.040 0.000 0.000 0.120 0.040 0.000 0.000 0.000 0.080 0.000 0.000 0.000 0.000
0.040 0.000 0.000 0.160 0.000 0.000 0.000 0.000
  left son=750284 (16 obs) right son=750285 (9 obs)
  Primary splits:
      Cough            < 3.035  to the right, improve=2.725556, (0 missing)
      Abdominal.obesity < 2.51   to the right, improve=2.548205, (0 missing)
      Fever            < 3.73   to the right, improve=2.520000, (0 missing)
```

```
      Constipation      < 3.2    to the left,  improve=2.336667, (0 missing)
      Acne              < 8.445  to the right, improve=2.253333, (0 missing)
  Surrogate splits:
      Fever        < 3.69   to the right, agree=0.92, adj=0.778, (0 split)
      Common.cold  < 6.21   to the right, agree=0.88, adj=0.667, (0 split)
      Sinusitis    < 1.295  to the right, agree=0.88, adj=0.667, (0 split)
      Constipation < 3.375  to the left,  agree=0.84, adj=0.556, (0 split)
      Infection    < 18.85  to the right, agree=0.80, adj=0.444, (0 split)


Node number 375163: 14 observations
  predicted class=Georgia             expected loss=0.7857143  P(node) =0.001062538
    class counts:    0    0    0    0    0    0    0    0    0    0    3    0    0    0    0    0    1    0
0    0    0    0    0    0    0    1    0    0    0    0    1    0    0    1    0    0    1    0    0    0    1
0    3    1    0    0    1    0    0    0    0
   probabilities: 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.214 0.000 0.000
0.000 0.000 0.000 0.071 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.071 0.000 0.000
0.000 0.000 0.071 0.000 0.000 0.071 0.000 0.000 0.071 0.000 0.000 0.000 0.071 0.000 0.214
0.071 0.000 0.000 0.071 0.000 0.000 0.000 0.000

Node number 375249: 30 observations,    complexity param=0.0002326303
  predicted class=Kansas              expected loss=0.8333333  P(node) =0.002276867
    class counts:    0    0    0    0    0    0    0    0    0    3    1    0    0    0    0    2    5    0
1    0    3    0    0    0    0    1    0    3    2    0    1    0    1    1    0    0    1    0    1    0    0
0    0    0    0    0    3    0    1    0    0
   probabilities: 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.100 0.033 0.000 0.000
0.000 0.000 0.067 0.167 0.000 0.033 0.000 0.100 0.000 0.000 0.000 0.000 0.033 0.000 0.100
0.067 0.000 0.033 0.000 0.033 0.033 0.000 0.000 0.033 0.000 0.033 0.000 0.000 0.000 0.000
0.000 0.000 0.000 0.100 0.000 0.033 0.000 0.000
  left son=750498 (7 obs) right son=750499 (23 obs)
  Primary splits:
      Nasal.congestion < 0.97   to the left,  improve=3.474534, (0 missing)
      Sinusitis        < 1.245  to the left,  improve=3.300621, (0 missing)
      Sore.throat      < 1.225  to the left,  improve=2.600000, (0 missing)
      Skin.condition   < 3.95   to the right, improve=2.500000, (0 missing)
      Abdominal.pain   < 3.965  to the left,  improve=2.333333, (0 missing)
  Surrogate splits:
      Cough            < 3.48   to the left,  agree=0.900, adj=0.571, (0 split)
      Sinusitis        < 1.22   to the left,  agree=0.900, adj=0.571, (0 split)
      Sore.throat      < 1.225  to the left,  agree=0.900, adj=0.571, (0 split)
      Insomnia         < 2.845  to the left,  agree=0.867, adj=0.429, (0 split)
      Autoimmune.disease < 1.22   to the left,  agree=0.833, adj=0.286, (0 split)

Node number 384151: 12 observations
```

```
  predicted class=New York             expected loss=0.5  P(node) =0.0009107468
    class counts:    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    1    0    0    0    0    0    0    0    0    0    3    0    6    0    0    0    0    0    0    0    0
0    0    1    0    0    1    0    0    0    0
   probabilities: 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.083 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.000 0.000 0.250 0.000 0.500 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.083 0.000 0.000 0.083 0.000 0.000 0.000 0.000

Node number 1505339: 23 observations,    complexity param=0.0001550868
  predicted class=West Virginia        expected loss=0.6086957  P(node) =0.001745598
    class counts:    0    0    0    2    0    0    0    0    0    0    0    0    0    0    0    0    0    3
0    0    0    0    0    0    0    4    0    0    1    0    0    0    0    1    0    1    1    0    0    0    0
0    1    0    0    0    0    0    9    0    0
   probabilities: 0.000 0.000 0.000 0.087 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.000 0.000 0.000 0.000 0.130 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.174 0.000 0.000
0.043 0.000 0.000 0.000 0.000 0.043 0.000 0.043 0.043 0.000 0.000 0.000 0.000 0.000 0.043
0.000 0.000 0.000 0.000 0.000 0.391 0.000 0.000
  left son=3010678 (15 obs) right son=3010679 (8 obs)
  Primary splits:
      Gastroesophageal.reflux.disease < 3.825  to the left,  improve=3.450000, (0 missing)
      Insomnia                        < 2.72   to the right, improve=2.523810, (0 missing)
      Sinusitis                       < 1.955  to the left,  improve=2.107143, (0 missing)
      Abdominal.obesity               < 1.955  to the right, improve=1.888889, (0 missing)
      Anxiety                         < 7.41   to the right, improve=1.875000, (0 missing)
  Surrogate splits:
      Abdominal.obesity < 1.865  to the right, agree=0.870, adj=0.625, (0 split)
      Abdominal.pain    < 3.855  to the right, agree=0.783, adj=0.375, (0 split)
      Arthritis         < 4.39   to the right, agree=0.783, adj=0.375, (0 split)
      Hypertension      < 6.535  to the left,  agree=0.783, adj=0.375, (0 split)
      Itch              < 6.04   to the right, agree=0.783, adj=0.375, (0 split)


Node number 3000937: 25 observations,    complexity param=0.0003101737
  predicted class=Georgia              expected loss=0.68  P(node) =0.001897389
    class counts:    0    0    0    4    0    0    0    0    0    0    8    0    0    0    0    0    1    0
2    0    0    0    0    0    0    0    0    0    0    0    4    0    0    0    0    0    4    0    0    0    0
0    0    1    0    0    0    0    1    0    0
   probabilities: 0.000 0.000 0.000 0.160 0.000 0.000 0.000 0.000 0.000 0.000 0.320 0.000 0.000
0.000 0.000 0.000 0.040 0.000 0.080 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.000 0.000 0.160 0.000 0.000 0.000 0.000 0.000 0.160 0.000 0.000 0.000 0.000 0.000 0.000
0.040 0.000 0.000 0.000 0.000 0.040 0.000 0.000
  left son=6001874 (11 obs) right son=6001875 (14 obs)
  Primary splits:
```

```
    Indigestion       < 1.44   to the right, improve=4.655584, (0 missing)
    Nausea            < 1.96   to the right, improve=3.547692, (0 missing)
    Back.pain         < 4.96   to the right, improve=3.509231, (0 missing)
    Infection         < 21.85  to the left,  improve=3.032208, (0 missing)
    Autoimmune.disease < 1.335  to the left,  improve=2.954286, (0 missing)
 Surrogate splits:
    Itch              < 8.49   to the right, agree=0.88, adj=0.727, (0 split)
    Autoimmune.disease < 1.335  to the left,  agree=0.84, adj=0.636, (0 split)
    Back.pain         < 4.96   to the right, agree=0.84, adj=0.636, (0 split)
    Constipation      < 3.23   to the right, agree=0.84, adj=0.636, (0 split)
    Infection         < 22.6   to the left,  agree=0.84, adj=0.636, (0 split)


> rsq.rpart(tree1)
Classification tree:
rpart(formula = state ~ . - date, data = training,    method = "class", control = rpart.control(cp =
0))

Variables actually used in tree construction:
 [1] Abdominal.obesity              Abdominal.pain                 Acne
Alcoholism
 [5] Allergy                        Anemia                         Anxiety
Arthritis
 [9] Asthma                         Attention.deficit.hyperactivity.disorder
Autoimmune.disease                  Back.pain
[13] Common.cold                    Constipation                   Cough
Depression
[17] Diabetes                       Diarrhea                       Fever
Gastroesophageal.reflux.disease
[21] Hypertension                   Indigestion                    Infection
Inflammation
[25] Insomnia                       Iron.deficiency                Itch
Migraine
[29] Nasal.congestion               Nausea                         Sinusitis
Skin.condition
[33] Skin.rash                      Sore.throat                    Stroke

Root node error: 12896/13176 = 0.97875

Summary
> summary(COVID19)
   state            Date          Abdominal.obesity Abdominal.pain      Acne         Alcoholism
Allergy
```

```
 Length:1098        Length:1098        Min.   :1.260     Min.   :3.140   Min.   : 6.480   Min.   : 3.000
Min.   : 6.88
 Class :character   Class :character   1st Qu.:2.070     1st Qu.:3.710   1st Qu.: 8.440   1st Qu.:
3.530   1st Qu.: 9.23
 Mode  :character   Mode  :character   Median :2.410     Median :3.960   Median : 9.210   Median
: 3.780   Median :10.09
                                       Mean   :2.534     Mean   :3.977   Mean   : 9.287   Mean   : 3.880   Mean
:10.17
                                       3rd Qu.:3.000     3rd Qu.:4.220   3rd Qu.: 9.930   3rd Qu.: 4.077   3rd
Qu.:11.05
                                       Max.   :4.160     Max.   :6.490   Max.   :13.860   Max.   :12.820   Max.
:15.94

    Anemia          Anxiety         Arthritis       Asthma          Attention.deficit.hyperactivity.disorder
 Min.   :1.250   Min.   : 5.590   Min.   : 3.200   Min.   :1.190   Min.   :1.610
 1st Qu.:1.810   1st Qu.: 7.140   1st Qu.: 4.240   1st Qu.:1.790   1st Qu.:2.400
 Median :2.050   Median : 7.690   Median : 4.490   Median :1.990   Median :2.670
 Mean   :2.037   Mean   : 7.744   Mean   : 4.767   Mean   :2.135   Mean   :2.740
 3rd Qu.:2.270   3rd Qu.: 8.270   3rd Qu.: 4.798   3rd Qu.:2.288   3rd Qu.:3.127
 Max.   :2.690   Max.   :15.260   Max.   :18.510   Max.   :5.730   Max.   :4.020

 Autoimmune.disease  Back.pain      Common.cold     Constipation    Cough          Depression
Diabetes
 Min.   :0.920       Min.   :3.33   Min.   : 4.520   Min.   :2.120   Min.   : 2.160   Min.   : 2.960   Min.
: 4.540
 1st Qu.:1.240       1st Qu.:4.72   1st Qu.: 6.213   1st Qu.:3.020   1st Qu.: 3.030   1st Qu.: 4.060
1st Qu.: 6.430
 Median :1.330       Median :5.07   Median : 7.730   Median :3.170   Median : 3.680   Median :
4.390   Median : 7.090
 Mean   :1.352       Mean   :5.03   Mean   :11.959   Mean   :3.161   Mean   : 5.150   Mean   : 4.453
Mean   : 7.087
 3rd Qu.:1.440       3rd Qu.:5.36   3rd Qu.:15.168   3rd Qu.:3.350   3rd Qu.: 7.438   3rd Qu.: 4.810
3rd Qu.: 7.750
 Max.   :2.390       Max.   :6.14   Max.   :54.320   Max.   :3.880   Max.   :16.920   Max.   :10.370
Max.   :15.710

    Diarrhea        Fever         Gastroesophageal.reflux.disease  Hypertension    Indigestion
Infection
 Min.   :2.680   Min.   : 2.780   Min.   :3.000                    Min.   :3.900   Min.   :1.070   Min.   :
13.52
 1st Qu.:3.280   1st Qu.: 3.580   1st Qu.:3.430                    1st Qu.:5.162   1st Qu.:1.320   1st Qu.:
18.29
 Median :3.595   Median : 4.210   Median :3.580                    Median :5.630   Median :1.400
Median : 20.02
```

```
 Mean   :3.614   Mean   : 5.068   Mean   :3.603                Mean   :5.628   Mean   :1.411   Mean
: 24.29
 3rd Qu.:3.938   3rd Qu.: 5.300   3rd Qu.:3.720                3rd Qu.:6.120   3rd Qu.:1.480   3rd
Qu.: 23.99
 Max.   :5.000   Max.   :22.380   Max.   :7.940                Max.   :7.210   Max.   :3.350   Max.
:100.00

  Inflammation      Insomnia      Iron.deficiency      Itch          Migraine      Nasal.congestion
Nausea
 Min.   :3.860   Min.   :1.870   Min.   :0.850   Min.   : 4.670   Min.   :2.900   Min.   :0.720   Min.
:1.400
 1st Qu.:4.793   1st Qu.:2.640   1st Qu.:1.280   1st Qu.: 5.760   1st Qu.:3.510   1st Qu.:1.090   1st
Qu.:1.660
 Median :5.060   Median :2.880   Median :1.470   Median : 6.555   Median :3.730   Median :1.320
Median :1.810
 Mean   :5.109   Mean   :2.891   Mean   :1.445   Mean   : 6.973   Mean   :3.756   Mean   :1.568
Mean   :1.825
 3rd Qu.:5.428   3rd Qu.:3.138   3rd Qu.:1.600   3rd Qu.: 8.185   3rd Qu.:3.950   3rd Qu.:2.100
3rd Qu.:1.980
 Max.   :7.700   Max.   :3.850   Max.   :2.120   Max.   :10.900   Max.   :6.660   Max.   :3.700
Max.   :2.720



  Sinusitis     Skin.condition     Skin.rash      Sore.throat       Stroke
 Min.   :0.890   Min.   :2.440   Min.   : 3.610   Min.   :0.880   Min.   :1.62
 1st Qu.:1.210   1st Qu.:3.442   1st Qu.: 5.130   1st Qu.:1.290   1st Qu.:2.25
 Median :1.370   Median :3.790   Median : 5.765   Median :1.420   Median :2.54
 Mean   :1.541   Mean   :3.746   Mean   : 6.038   Mean   :1.775   Mean   :2.62
 3rd Qu.:1.917   3rd Qu.:4.080   3rd Qu.: 6.700   3rd Qu.:2.268   3rd Qu.:2.96
 Max.   :2.890   Max.   :4.600   Max.   :25.650   Max.   :5.090   Max.   :6.15



'data.frame':       1098 obs. of  37 variables:
 $ state                          : chr  "California" "California" "California" "California" ...
 $ Date                           : chr  "1/1/2020" "1/2/2020" "1/3/2020" "1/4/2020" ...
 $ Abdominal.obesity                    : num  2.39 2.52 2.29 2.39 2.59 2.58 2.53 2.51 2.41 2.05 ...
 $ Abdominal.pain                       : num  4.8 4.74 4.65 4.75 4.64 4.56 4.47 4.36 4.47 4.22 ...
 $ Acne                           : num  10.2 11.2 10.9 11.3 11.4 ...
 $ Alcoholism                        : num  6.57 5.05 4.79 4.92 4.77 4.38 4.25 4.35 4.55 4.57 ...
 $ Allergy                        : num  9.62 10.4 10.42 10.14 9.75 ...
 $ Anemia                         : num  1.61 2.03 2.05 1.9 1.68 2.12 2.3 2.38 2.36 2.24 ...
 $ Anxiety                        : num  7.3 7.93 7.57 7.32 7.69 8.12 8.18 8.26 8.25 7.55 ...
 $ Arthritis                      : num  4.41 4.87 4.77 4.76 4.58 4.72 4.77 4.83 4.89 4.52 ...
 $ Asthma                         : num  1.67 2.2 2.34 1.94 1.82 2.19 2.32 2.22 2.23 2.1 ...
```

```
$ Attention.deficit.hyperactivity.disorder: num  2.08 2.5 2.56 2.28 2.27 2.77 3.02 3.69 3.39 2.97
...
$ Autoimmune.disease                 : num  1.26 1.33 1.28 1.2 1.19 1.34 1.37 1.55 1.47 1.39 ...
$ Back.pain                          : num  5.26 6.02 5.75 5.7 5.68 5.99 5.96 5.81 5.71 5.28 ...
$ Common.cold                        : num  19 19.8 19.9 20.1 19 ...
$ Constipation                       : num  3.29 3.21 3.15 3.41 3.38 3.1 3.19 3.27 3.28 3.16 ...
$ Cough                              : num  10.9 11.6 11.7 11.6 11.1 ...
$ Depression                         : num  4.71 4.75 4.5 4.52 4.71 4.82 4.91 4.96 5.03 4.63 ...
$ Diabetes                           : num  6.2 7.41 7.38 6.77 6.79 7.9 8.18 8.29 8.45 7.92 ...
$ Diarrhea                           : num  4.46 4.36 4.31 4.41 4.34 4.19 4.04 4.09 4.15 3.98 ...
$ Fever                              : num  5.41 5.51 5.59 5.69 5.41 5.39 5.38 5.22 5.09 5.08 ...
$ Gastroesophageal.reflux.disease    : num  3.89 4.02 3.9 3.86 3.78 3.84 3.77 3.7 3.73 3.8 ...
$ Hypertension                       : num  4.76 6.33 6.14 5.37 5.22 6.32 6.4 6.42 6.58 6.27 ...
$ Indigestion                        : num  1.93 1.72 1.69 1.73 1.7 1.64 1.55 1.49 1.5 1.51 ...
$ Infection                          : num  20.2 22.5 23.1 22 21.1 ...
$ Inflammation                       : num  5.39 6.12 6 5.89 5.74 6.07 6.21 6.1 6.08 5.8 ...
$ Insomnia                           : num  2.85 3.38 3.34 3.21 3.51 3.85 3.77 3.7 3.55 3.3 ...
$ Iron.deficiency                    : num  1.29 1.57 1.59 1.51 1.37 1.62 1.69 1.73 1.74 1.63 ...
$ Itch                               : num  6.64 6.82 6.85 6.96 6.91 6.77 6.71 6.54 6.48 6.19 ...
$ Migraine                           : num  3.95 4.07 4.05 4.03 3.87 4.21 4.2 4.1 4.03 3.99 ...
$ Nasal.congestion                   : num  3.7 3.6 3.63 3.69 3.48 3.22 3.28 3.19 2.96 2.82 ...
$ Nausea                             : num  2.72 2.53 2.45 2.43 2.36 2.36 2.25 2.2 2.16 2.1 ...
$ Sinusitis                          : num  2.49 2.58 2.63 2.6 2.49 2.46 2.49 2.39 2.27 2.16 ...
$ Skin.condition                     : num  3.6 4.07 4.05 3.88 3.75 3.99 4.03 4.03 4.03 3.79 ...
$ Skin.rash                          : num  5.77 5.94 5.9 5.86 5.73 ...
$ Sore.throat                        : num  3.32 3.47 3.47 3.43 3.23 3.13 3.12 2.98 2.87 2.7 ...
$ Stroke                             : num  3.26 2.97 2.84 2.51 2.41 3.19 3.19 3.3 3.4 3.02 .
```

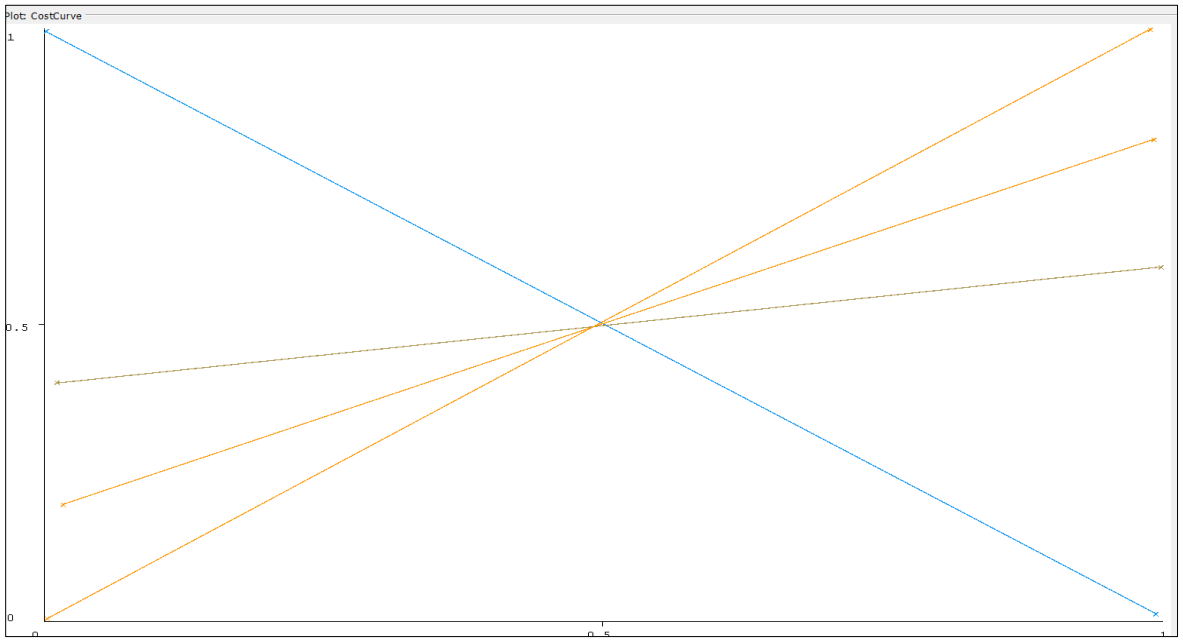
REP TREE ANALYSIS RESULT

```
REPTree
============
 : California (732/488) [366/244]

Size of the tree : 1

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         360               32.7869 %
Incorrectly Classified Instances       738               67.2131 %
Kappa statistic                         -0.0082
Mean absolute error                      0.4444
```

```
Root mean squared error                  0.4714
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1098

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.393    0.402    0.329      0.393    0.358      -0.008   0.496     0.331     California
               0.197    0.202    0.327      0.197    0.246      -0.006   0.496     0.331     New Jersey
               0.393    0.404    0.327      0.393    0.357      -0.011   0.494     0.330     New York
Weighted Avg.  0.328    0.336    0.328      0.328    0.320      -0.008   0.495     0.331

=== Confusion Matrix ===

   a   b   c   <-- classified as
 144  74 148 |   a = California
 146  72 148 |   b = New Jersey
 148  74 144 |   c = New York

Classification Accuracy is 32.79%
```

Experimented with decision tree analysis the above shows the model doesn't fit for this type of data as the confidence accuracy is pretty low.

Cost Curve for California state



X – Probability Cost Function

Y- Normalized Expected Cost

EXPERIMENT WITH NAÏVE BAYES

Naïve bayes is one among the important to analyze the data of such types here with the kind of data we have and probability of having impact of virus due to existing health conditions, even though results are that great, its far better than the decision tree models.
Provided with an accuracy of "64.39%" which is a good indicator to draw predictions for the key states NY.NJ and CA for the three quarters that is people who are having some existing health conditions have highly impacted with virsus.

```
=== Summary ===

Correctly Classified Instances         707               64.3898 %
Incorrectly Classified Instances       391               35.6102 %
Kappa statistic                          0.4658
K&B Relative Info Score                 49.8338 %
K&B Information Score                  867.267  bits      0.7899 bits/instance
Class complexity | order 0            1740.3195 bits      1.585  bits/instance
Class complexity | scheme             4136.6723 bits      3.7675 bits/instance
Complexity improvement     (Sf)      -2396.3527 bits     -2.1825 bits/instance
Mean absolute error                      0.2475
Root mean squared error                  0.4482
Relative absolute error                 55.6947 %
Root relative squared error             95.0836 %
Total Number of Instances               1098

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.790    0.150    0.724      0.790    0.756      0.627    0.886     0.833     California
               0.456    0.175    0.566      0.456    0.505      0.299    0.712     0.518     New Jersey
               0.686    0.209    0.621      0.686    0.652      0.466    0.807     0.627     New York
Weighted Avg.  0.644    0.178    0.637      0.644    0.638      0.464    0.802     0.660

=== Confusion Matrix ===

   a   b   c   <-- classified as
 289  47  30 |   a = California
  76 167 123 |   b = New Jersey
  34  81 251 |   c = New York
```

Here the accuracy shows a better result than that of decision tree

Naïve Bayes provides a sort of conclusion that there are high chances of the getting influenced to the virus who has the underlined issues with Asthma, Allergy, Sinusitis, Cough, Cold, Sore throat and Stroke.

EXPERIMENT WITH RANDOM FOREST METHOD

Likewise experiment with Random Forest also provided a better accuracy

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         660               60.1093 %
Incorrectly Classified Instances       438               39.8907 %
Kappa statistic                          0.4016
K&B Relative Info Score                 28.1305 %
K&B Information Score                  489.5599 bits      0.4459 bits/instance
Class complexity | order 0            1740.3195 bits      1.585  bits/instance
Class complexity | scheme             1310.7981 bits      1.1938 bits/instance
Complexity improvement     (Sf)        429.5215 bits      0.3912 bits/instance
Mean absolute error                      0.3566
Root mean squared error                  0.408
Relative absolute error                 80.2394 %
Root relative squared error             86.5389 %
Total Number of Instances               1098

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.869     0.046     0.903      0.869    0.886      0.831    0.970     0.961     California
               0.470     0.321     0.423      0.470    0.445      0.145    0.565     0.366     New Jersey
               0.464     0.231     0.501      0.464    0.482      0.238    0.776     0.506     New York
Weighted Avg.  0.601     0.199     0.609      0.601    0.604      0.405    0.770     0.611

=== Confusion Matrix ===

   a   b   c   <-- classified as
 318  41   7 |   a = California
  32 172 162 |   b = New Jersey
   2 194 170 |   c = New York

Accuracy as 60.19%
```

Conclusion :

The above accuracy is also satisfactory however the output did not provide a greater confidence when compared with all variables, however this result also provides the forecast and trend as i.e, people who have some issues like Asthma, Sinusitis and cold, fever, nasal conjunction have more impacted with virus.

EXPERIMENT WITH SUPPORT VECTOR MACHINES

Choose the kernel option as “poly kernel” and set the random seed as “1” figured out the resultant as below with the highest accuracy as 91.07%

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1000               91.0747 %
Incorrectly Classified Instances        98                8.9253 %
Kappa statistic                          0.8661
Mean absolute error                      0.2443
Root mean squared error                  0.3091
Relative absolute error                 54.9631 %
Root relative squared error             65.564  %
Total Number of Instances             1098

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.959    0.011    0.978      0.959    0.968      0.953    0.990     0.982     California
               0.863    0.042    0.911      0.863    0.886      0.832    0.928     0.841     New Jersey
               0.910    0.081    0.849      0.910    0.879      0.816    0.923     0.808     New York
Weighted Avg.  0.911    0.045    0.913      0.911    0.911      0.867    0.947     0.877

=== Confusion Matrix ===

   a   b   c   <-- classified as
 351   2  13 |   a = California
   4 316  46 |   b = New Jersey
   4  29 333 |   c = New York
```

Conclusion

With the SVM experiment we have accomplished a higher confidence level than rest of the techniques i.e.., “91.07%”, and the classification shows the impact of existing conditions influenced people of the three states that are considered for the three quarters.

Their precision values are also high to draw this conclusion.

Among the experimented methods below methods did provide a satisfactory result, however those confidence levels are not as great as anticipated

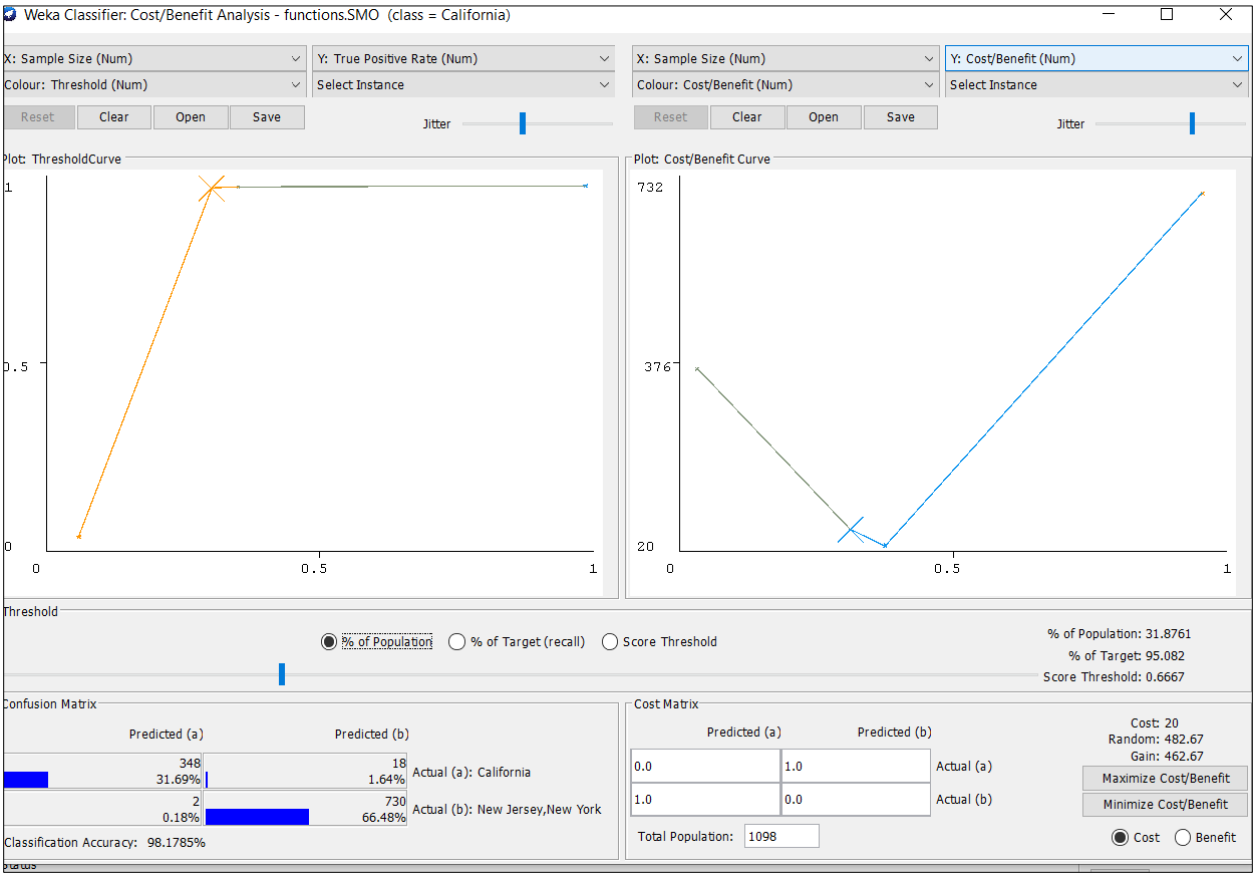
1. Decision Tree Techniques
2. Naïve Bayes classifiers
3. Knn Models

The technique that provided a good result and highest accuracy is “Support Vector Machines” with a good confidence level and accuracy “91.07”, to provide an indicator to the various systems that people with certain existing conditions are impacted highly by the virus.

This conclusion drawn for three key states for the first three quarters, if we process entire data set across all states we can accomplish the similar results.

Even though it’s a research data, just providing trends, forecast will alone be beneficial, if we can give a slight information how economically it will have impact over a period of time on various aspects of improving health conditions.

Unfortunately, the data set I’ve chosen does not contain much attributes to calculate the same however with the help of results set predicted the cost benefit for California state



REFERENCES

1. https://cran.r-project.org/web/packages/covid19.analytics/vignettes/covid19.analytics.html
2. https://pair-code.github.io/covid19_symptom_dataset/?country=US&symptom=asthma
3. https://packagemanager.rstudio.com/client/#/repos/1/packages/gensvm
4. https://www.r-graph-gallery.com/histogram_several_group.html
5. https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/
6. https://coronavirus.jhu.edu/data/new-cases
7. https://www.sciencedirect.com/science/article/pii/S235234092031249X
8. https://www.ijert.org/prediction-and-analysis-of-data-mining-models-for-students-underlying-issues-during-novel-coronavirus-covid-19
9. https://www.analyticsvidhya.com/blog/2020/03/decision-tree-weka-no-coding/
10.