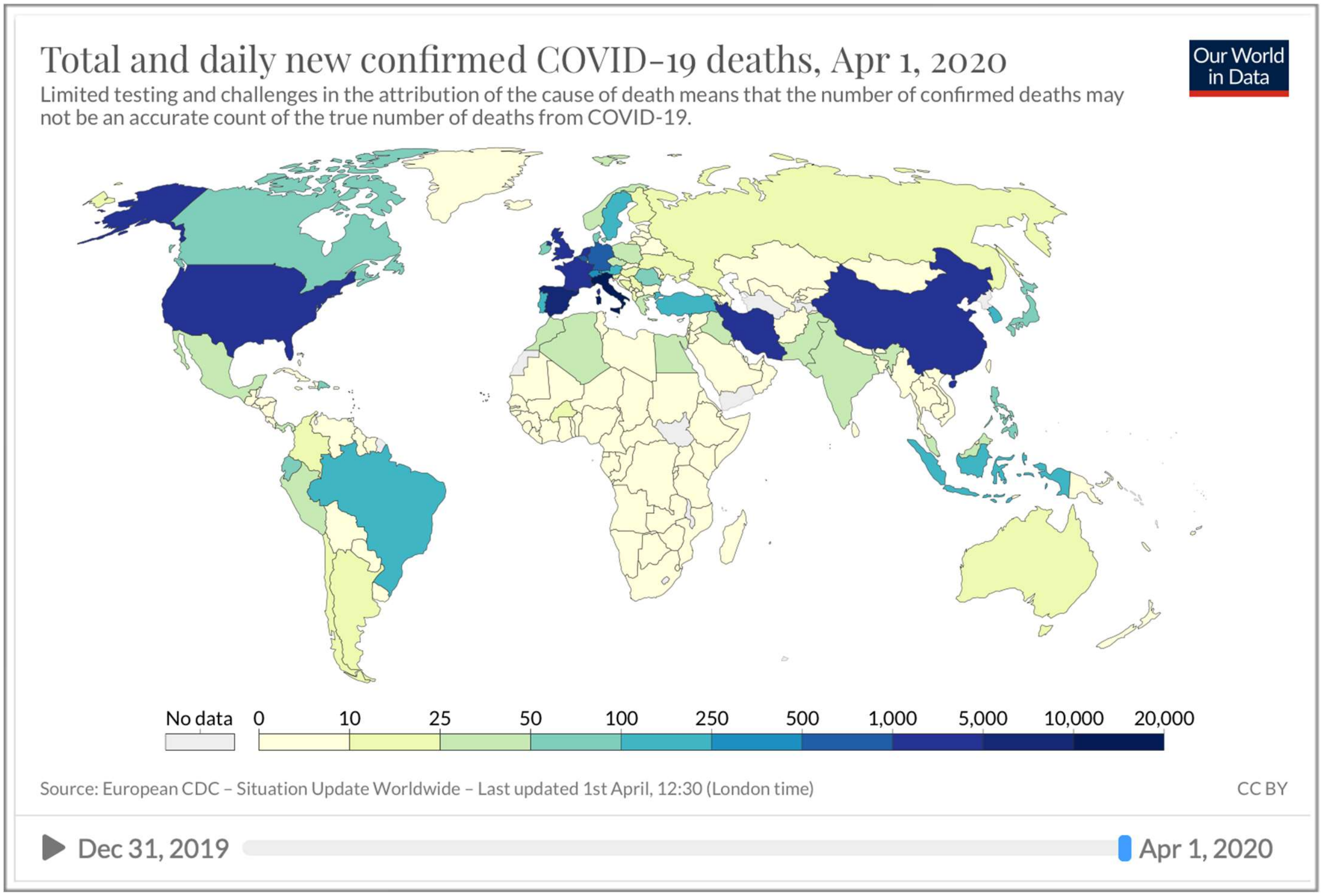
A Data Mining & Machine learning Project

SASTRA DEEMED TO BE UNIVERSITY, THANJAVUR

Done by:
R.Nandhakumar- 122013026
R.Shashank-122013041
T.M.N.Yatindrapravanan- 122013059

A Project on COVID-19 - Identifying the Hotspot States in India in a comparative analysis with other states



Total and daily new confirmed COVID-19 deaths, Apr 1, 2020

Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19.

Source: European CDC – Situation Update Worldwide – Last updated 1st April, 12:30 (London time)

Table of Contents

Introduction

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19 - World Health Organization. The number of new cases are increasing day by day around the world.   - WHO

People can catch COVID-19 from others who have the virus. This has been spreading rapidly around the world and Italy is one of the most affected country.

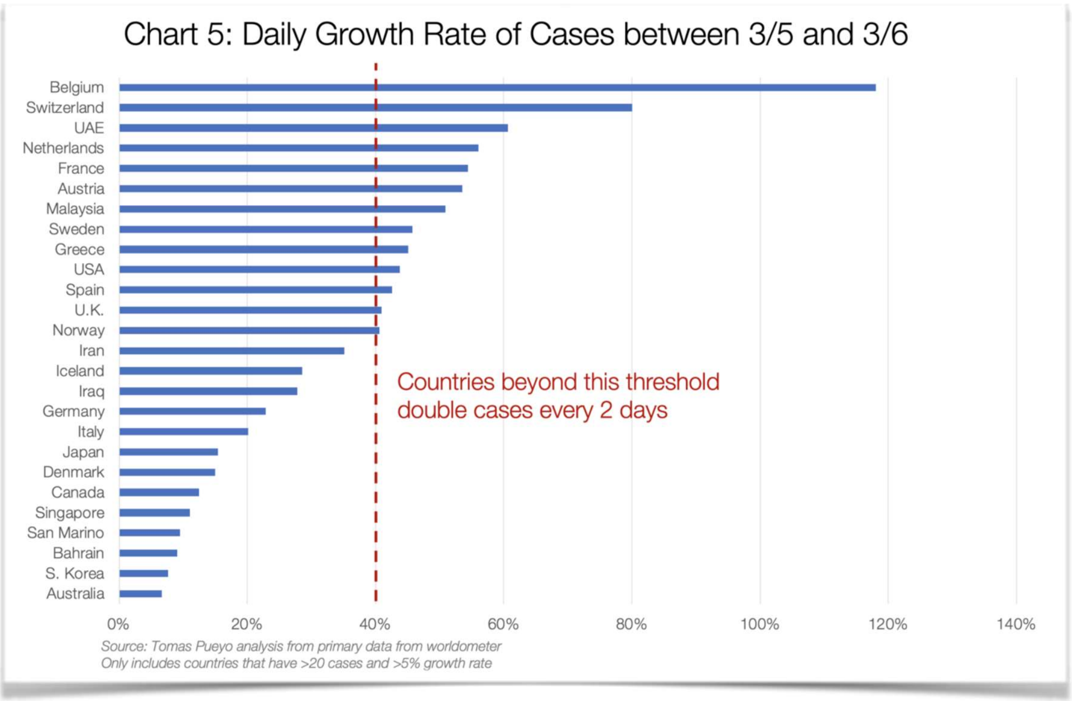


Chart 5: Daily Growth Rate of Cases between 3/5 and 3/6

Source: Tomas Pueyo analysis from primary data from worldometer
Only includes countries that have >20 cases and >5% growth rate

So, it has to be noted that the growth rate once if cross the 40% threshold, the situation gets worse significantly for a nation. It can be noted from various sources like the NBC News that Northern Italy has one of the best public health systems in the Western world. Its doctors and medical professionals are well-trained. They felt prepared when the coronavirus began to spread through their prosperous, well-educated region. And they still could do nothing to prevent what happened.

Italy will quarantine the entire Lombardy region around Milan to limit the spread of the coronavirus as well as areas around and including Venice and the cities of Parma and Rimini, Italian media reported Saturday.

- The Economic Times, March 8, 2020

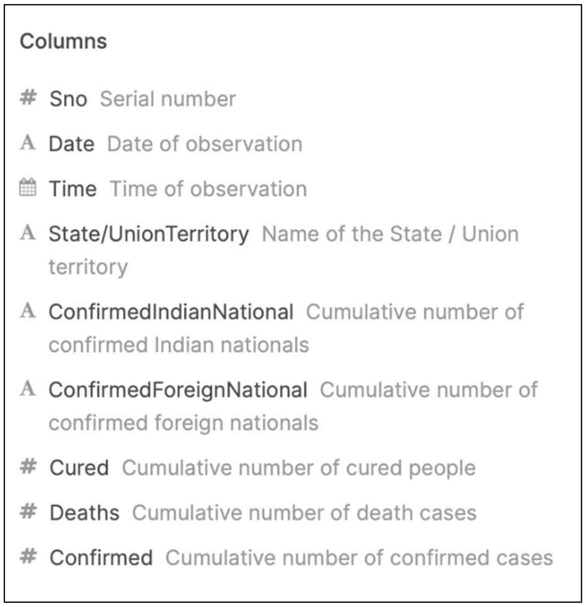
The same article also told, *the government decree also covers parts of the Veneto region around Venice as well as Emilia-Romagna's Parma and Rimini*. So, taking these 2 regions- Veneto and Lombardy, it was observed from the growth rate data, that it was nearly 22-23% on march 8th. And by that time, many people were infected which is evident for us to say that it was late for this quarantine to be imposed.

Similarly, in India, our target of study, has shown much related data between a day and the previous. Significant rises in particular regions followed a similar or more favorable pattern to that of in Italy.

Introduction to the dataset

The dataset considered for this study, is based on COVID-19 cases at daily level present in India from kaggle datasets.
Thanks to Indian Ministry of Health & Family Welfare for making the data available to general public.
Dataset considered is from 30/01/2020 to 02/04/2020.

❖ Description of attributes in the dataset:



Columns

#	Sno	Serial number
A	Date	Date of observation
📅	Time	Time of observation
A	State/UnionTerritory	Name of the State / Union territory
A	ConfirmedIndianNational	Cumulative number of confirmed Indian nationals
A	ConfirmedForeignNational	Cumulative number of confirmed foreign nationals
#	Cured	Cumulative number of cured people
#	Deaths	Cumulative number of death cases
#	Confirmed	Cumulative number of confirmed cases

The data preprocessing required to finalize the reliable data columns amongst of all these abundant data. The factors taken concern of are:

- Locating the case
- Future of the region
- The relation between today and tomorrow of the region
- Cumulative count of cases in the region.

All of these were taken care of and brought to a conclusion on the data columns required and the needed was done.

Objectives

- To identify, classify and inform of potential hotspots of COVID 19 in India
  - The model classifies states into various categories - Hotspots, Potential Hotspots, Manageable and COVID free. This classification is based on appropriate data analysis which will be detailed in the paper.
- To provide data and analysis about said hotspots to aid in further decision making and policies.
- To identify potential hotspots to help in prevention and in related policy decisions
- To perform analysis with accuracy, inform of any error and include possible error adjustment/correction methods to arrive at appropriate conclusions.
- This model also strives to be a hotspot detection model for any such future pandemics of a virulent nature.

Methods

- ## Data Preprocessing

A. Handling NULL values or '-'s :

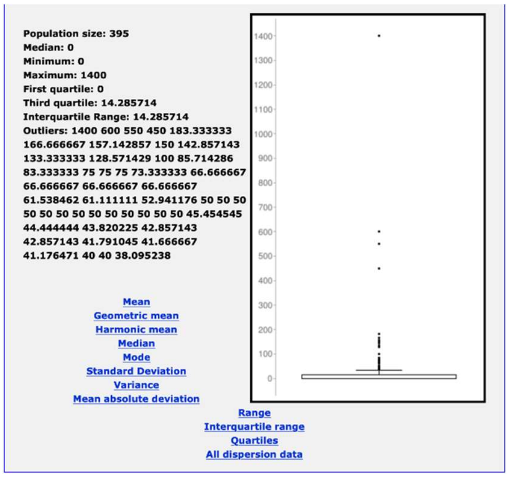
The '-'s present in the dataset were replaced with 0 which means for the initial days the number of COVID-19 confirmed cases in a State/UT was 0. Here, the null values can not be replaced with the mean value as it will change the meaning of the # Confirmed attribute.

B. Nominal to date conversion of an attribute :

The 'Date' attribute was a nominal attribute which was changed to Date type as it provides the required meaning. For this, a php script was used as the format for the attribute to be Date type was not followed in the dataset.

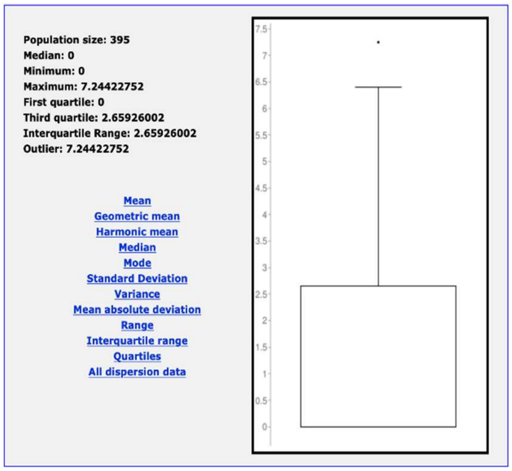
C. The log transformation:

The data after the first level training resulted in hugely ranged set. This might not be a valid training for a right prediction. Therefore to increase the prediction rate more accurate and for a better test data validation, we have done the log transformation to the increase rate column data. This resulted in a better ranged data column to carry on with prediction.



Population size: 395
Median: 0
Minimum: 0
Maximum: 1400
First quartile: 0
Third quartile: 14.285714
Interquartile Range: 14.285714
Outliers: 1400 600 550 450 183.333333 166.666667 157.142857 150 142.857143 133.333333 128.571429 100 85.714286 83.333333 75 75 75 73.333333 66.666667 66.666667 66.666667 66.666667 61.538462 61.111111 52.941176 50 50 50 50 50 50 50 50 50 50 50 50 50 45.454545 44.444444 43.820225 42.857143 42.857143 41.791045 41.666667 41.176471 40 40 38.095238

Figure 1 before log transformation



Population size: 395
Median: 0
Minimum: 0
Maximum: 7.24422752
First quartile: 0
Third quartile: 2.65926002
Interquartile Range: 2.65926002
Outlier: 7.24422752

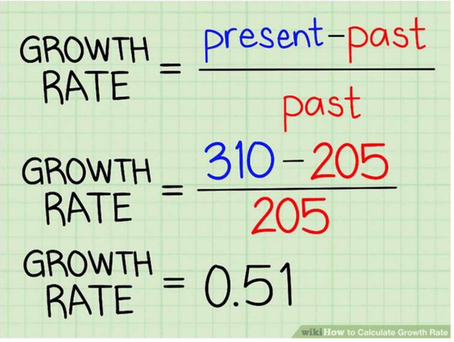
Figure 2 after log transformation

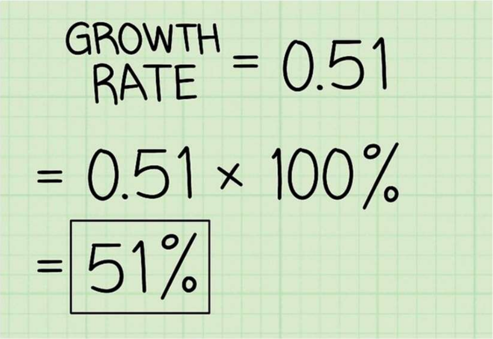
D. Removing redundant attributes :

For this study, the attributes such as ConfirmedIndianNational, ConfirmedForeignNational, Cured and Deaths are not required as the aim is only to study the Confirmed cases and the growth rate(or the increase rate), which are the measures reliable to study the ability of the virus to spread in a specific state, are to be considered.

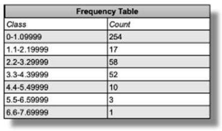
E. Adding necessary prospects in the dataset :

As stated above, the increase rate needs to be considered. Growth rate is calculated using the formula given below.

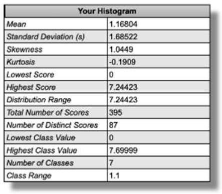

GROWTH RATE = present-past / past
GROWTH RATE = 310-205 / 205
GROWTH RATE = 0.51
wikiHow to Calculate Growth Rate



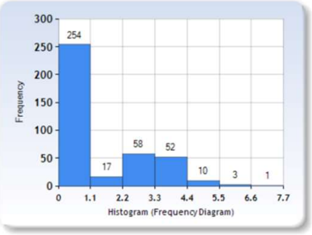
Dataset validation:

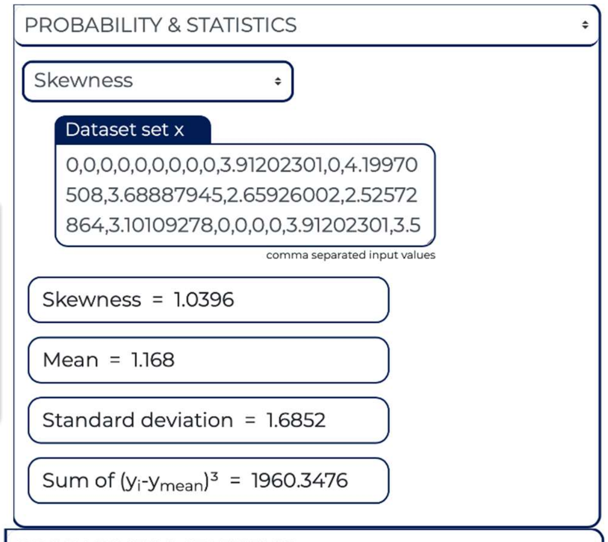


Frequency Table	
Class	Count
0-1.09999	254
1.1-2.19999	17
2.2-3.29999	58
3.3-4.39999	52
4.4-5.49999	10
5.5-6.59999	3
6.6-7.69999	1



Your Histogram	
Mean	1.16804
Standard Deviation (s)	1.68522
Skewness	1.0449
Kurtosis	-0.1909
Lowest Score	0
Highest Score	7.24423
Distribution Range	7.24423
Total Number of Scores	395
Number of Distinct Scores	87
Lowest Class Value	0
Highest Class Value	7.69999
Number of Classes	7
Class Range	1.1





Dataset set x

0,0,0,0,0,0,0,0,0,3.91202301,0,4.19970508,3.68887945,2.65926002,2.52572864,3.10109278,0,0,0,0,3.91202301,3.5

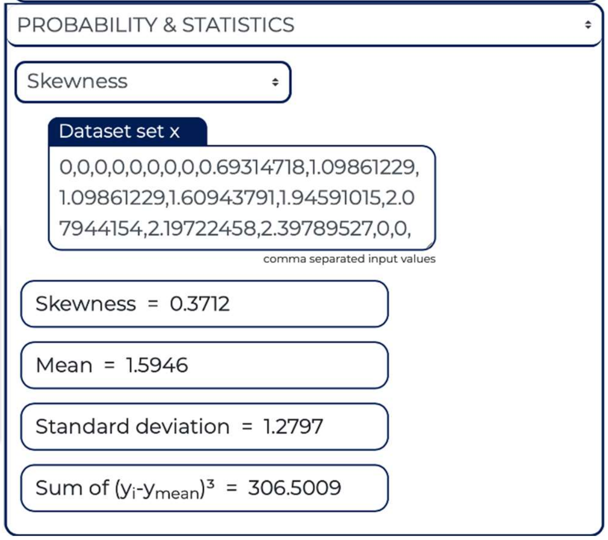
comma separated input values

Skewness = 1.0396

Mean = 1.168

Standard deviation = 1.6852

Sum of $(y_i\text{-}y_{mean})^3$ = 1960.3476



Dataset set x

0,0,0,0,0,0,0,0,0.69314718,1.09861229,1.09861229,1.60943791,1.94591015,2.07944154,2.19722458,2.39789527,0,0,

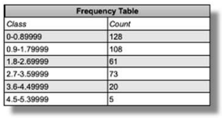
comma separated input values

Skewness = 0.3712

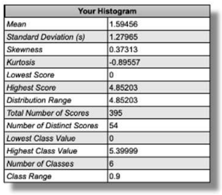
Mean = 1.5946

Standard deviation = 1.2797

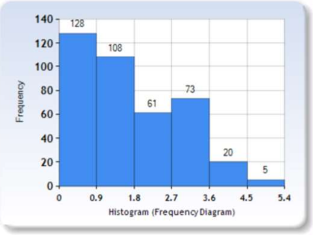
Sum of $(y_i\text{-}y_{mean})^3$ = 306.5009



Frequency Table	
Class	Count
0-0.89999	128
0.9-1.79999	108
1.8-2.69999	61
2.7-3.59999	73
3.6-4.49999	20
4.5-5.39999	5



Your Histogram	
Mean	1.59456
Standard Deviation (s)	1.27965
Skewness	0.37313
Kurtosis	-0.89557
Lowest Score	0
Highest Score	4.85203
Distribution Range	4.85203
Total Number of Scores	395
Number of Distinct Scores	54
Lowest Class Value	0
Highest Class Value	5.39999
Number of Classes	6
Class Range	0.9



Methods

The growth rate was calculated for every day for every state and this was used as an attribute. This attribute was used to classify a state as a Hotspot or not. But the problem with this measure is that, for a State/UT- the growth rate of the initial period i.e the growth rate between the day the state/UT reported the first case and the next day will be very high as the denominator will contain just 1 which means the numerator*100 times, which will be a huge number and if the machine learning model predicts using this sort of a data, it will be not a valid output as classifying a state as a hotspot in the first day itself, does not make sense in real life.

Another issue with this measure is, there can be recovered people so the number of positive cases will vary the next day, which will give a negative growth rate value which lowers the preference of this state to be a hotspot while it is not.

So, to solve both the cases, the attribute is renamed as increase_rate_preference and the data values of the above discussed cases were set to 0, which does not change the meaning of the data.

The dataset now has description as given,

1. Date - DATE
2. Time - NOMINAL
3. State/Union Territory - NOMINAL
4. Total_Positive_Cases - NUMERIC
5. Increase_rate_preference - NUMERIC
6. Hotspot_Prediction - BINARY

This data was now separated as training set(before 26 march) and test set(after 26 march) as the outcome variable for that time period is known from the WHO report about COVID-19-India on 22-March-2020 as- Maharashtra, Kerala, Delhi and UP had the most number of cases of COVID-19. It can also be observed from the dataset that the growth rate was maximum for these States before 26th March which means these states have positive cases which spread the virus rapidly.

Model validation:

Various classifier models have been tried and validated for their efficiency to our prediction and a better result rearing model is chosen for prediction, which is the SMO (Support Vector Machine)classifier model.

S No.	Algorithm	Accuracy Obtained
1	J48 decision tree	89.1753 %
2	JRip	90.2991 %
3	Support Vector Machine (SMO in Weka)	89.5696 %
4	Naïve Bayes	78.8660 %
5	KStar	89.5332 %

Support Vector Machines algorithm was selected for building the final model even though "JRip" algorithm has a higher accuracy rate.

This was because:

i) The accuracies of the algorithms are very close to each other. So, either of them can be used for the model

ii) On comparing the predictions done by both the algorithms, SMO's predictions were closer to the statistically classified real time data.

Observations

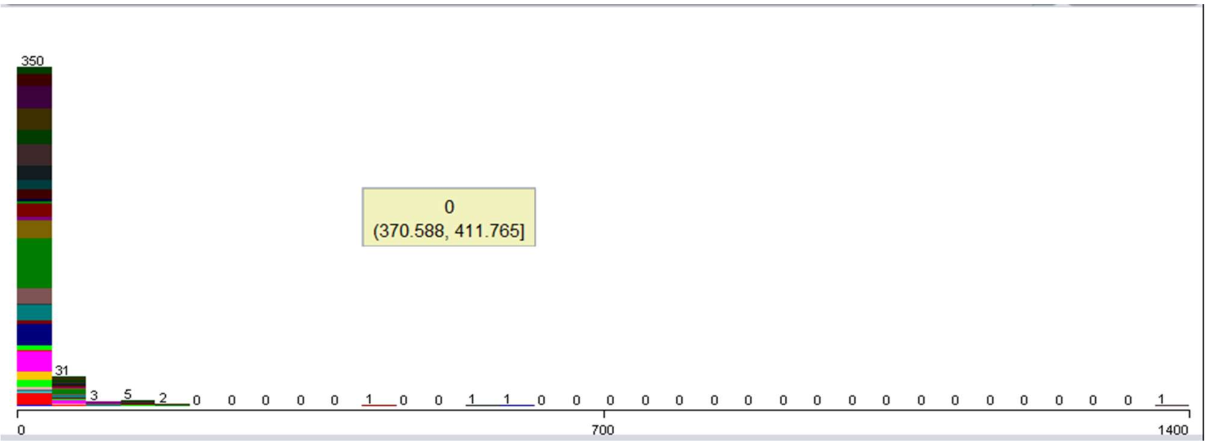
Training data observations:

The training data has been observed for various analysis on the WEKA explorer platform. It is considered as a measure to ensure the data's nature is well suited for all the further proceedings.

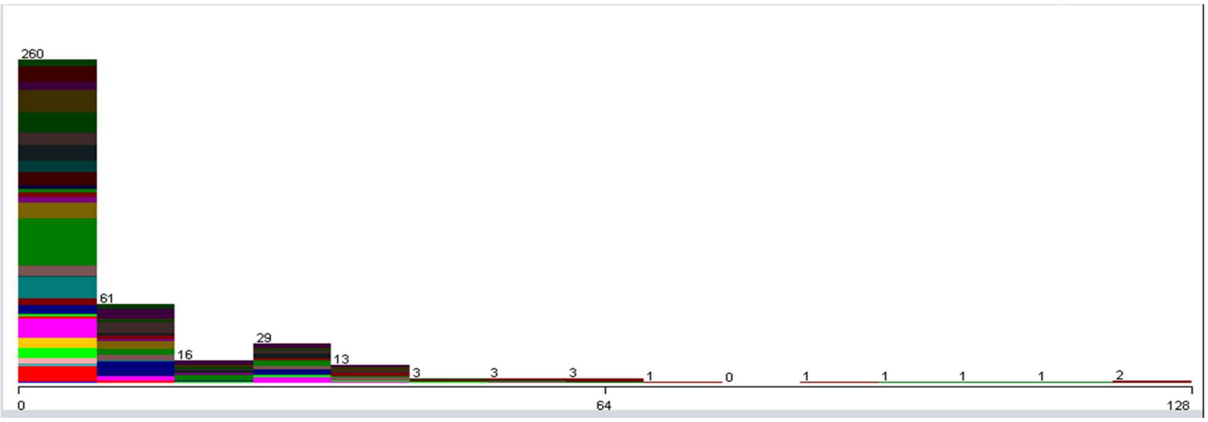
The entire dataset was split into train, test and predict accordingly to obtain the optimum results out of it.

- Till March 26, 2020 - training data
- Till April 2, 2020 - test data
- Till April 10, 2020 - prediction data

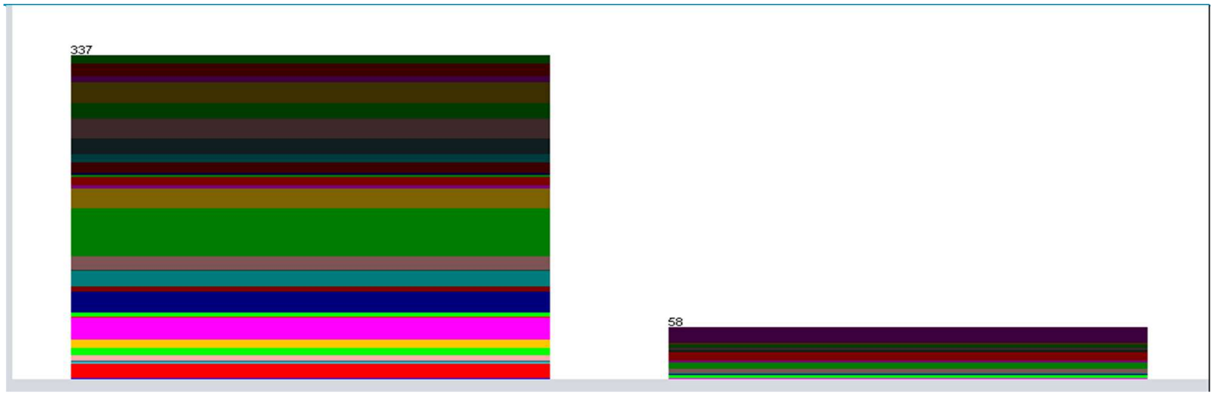
Given the observations of the training data set:



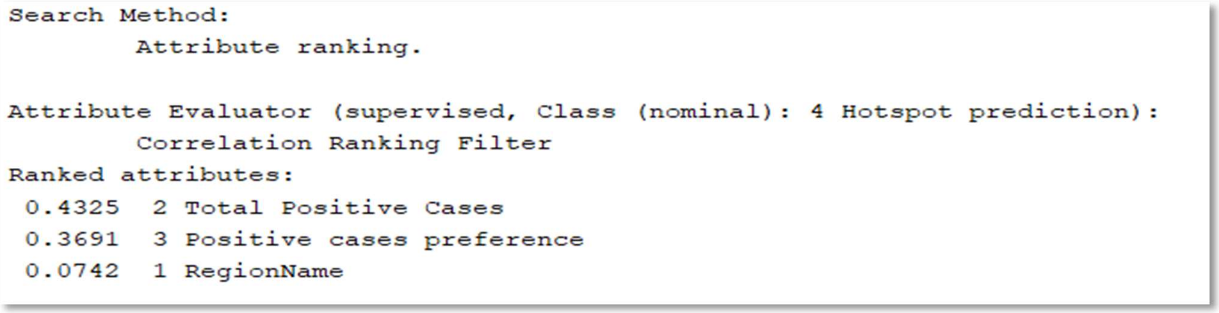
graph i: the increase rate (class: states)



graph ii: the total positive cases

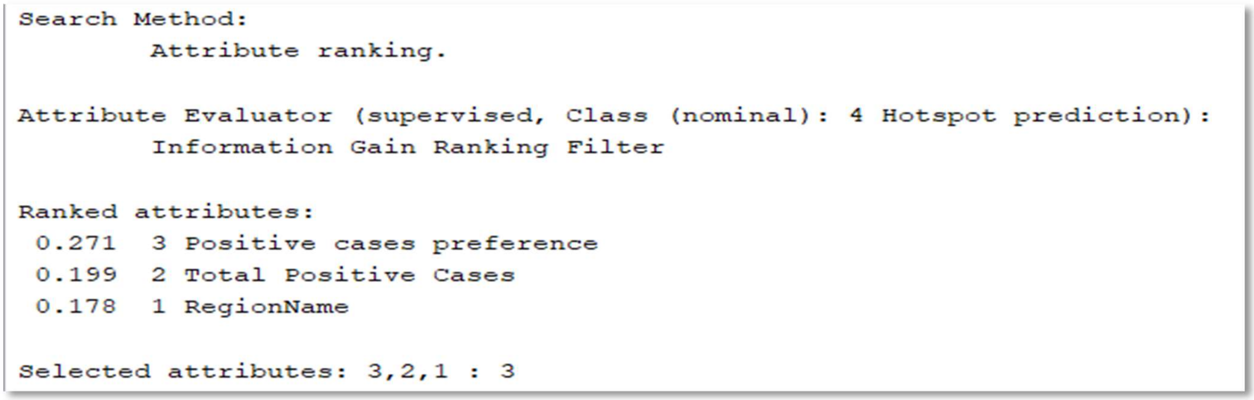


graph iii: potential hotspots



```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 4 Hotspot prediction):
        Correlation Ranking Filter
Ranked attributes:
 0.4325   2 Total Positive Cases
 0.3691   3 Positive cases preference
 0.0742   1 RegionName
```

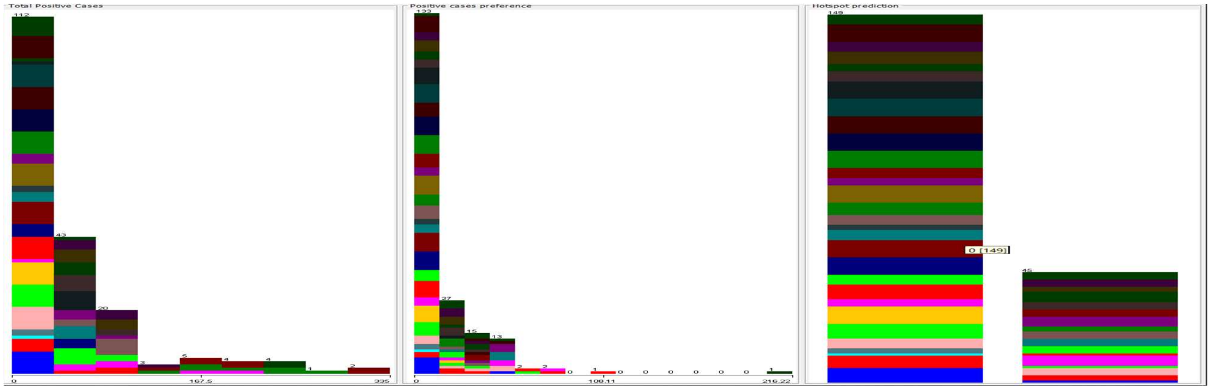
info 1: correlation of training data set



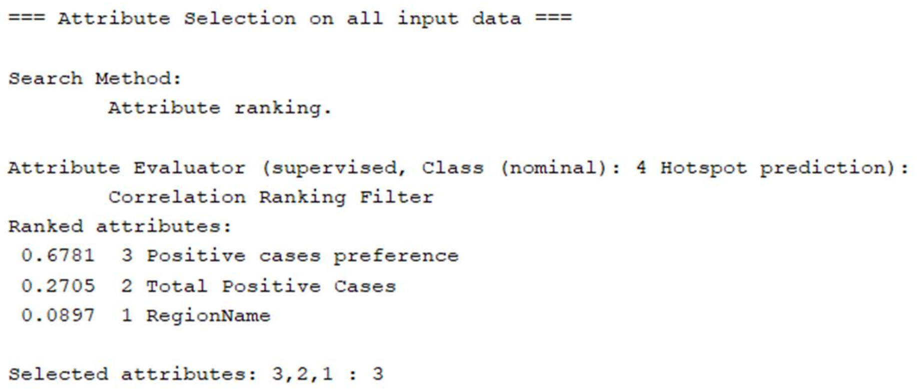
```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 4 Hotspot prediction):
        Information Gain Ranking Filter

Ranked attributes:
 0.271   3 Positive cases preference
 0.199   2 Total Positive Cases
 0.178   1 RegionName

Selected attributes: 3,2,1 : 3
```

info 2: info gain of the training data set

test data set observations:

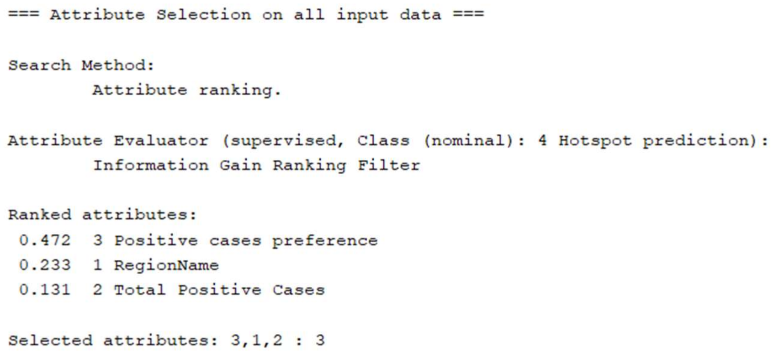


graph iv test data set observations



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 4 Hotspot prediction):
        Correlation Ranking Filter
Ranked attributes:
 0.6781   3 Positive cases preference
 0.2705   2 Total Positive Cases
 0.0897   1 RegionName

Selected attributes: 3,2,1 : 3
```

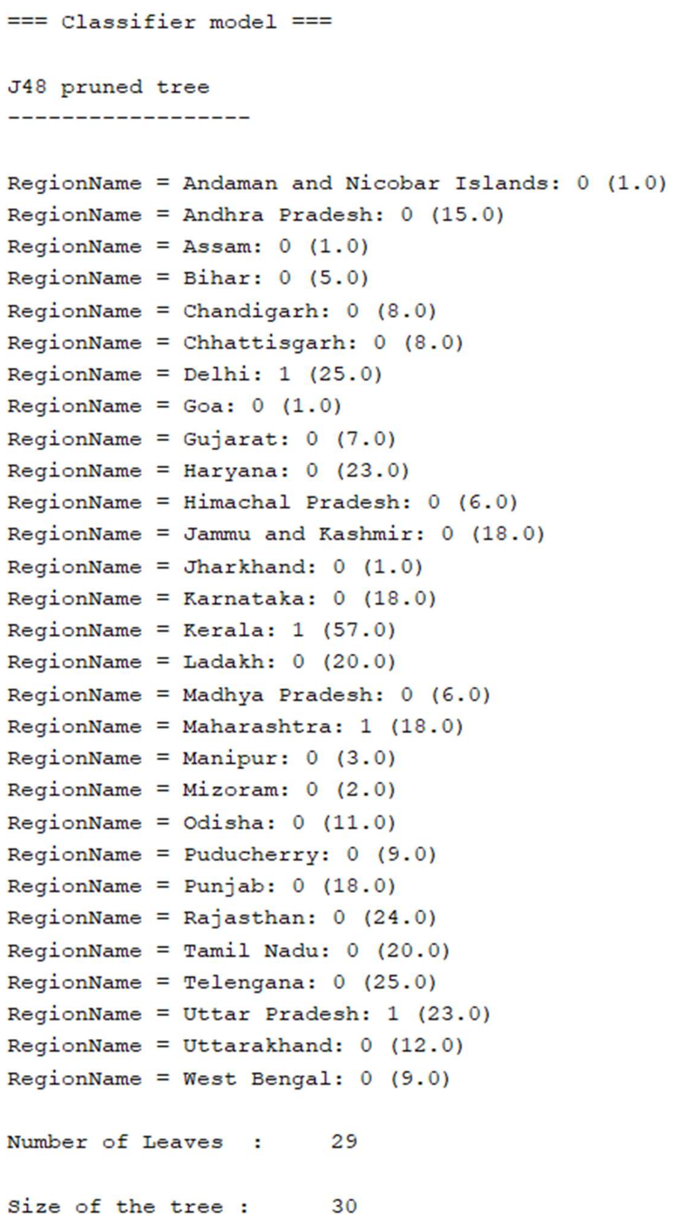
info 3 test data corelation



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 4 Hotspot prediction):
        Information Gain Ranking Filter

Ranked attributes:
 0.472   3 Positive cases preference
 0.233   1 RegionName
 0.131   2 Total Positive Cases

Selected attributes: 3,1,2 : 3
```

info 4 test data information gain

prediction results:



```
=== Classifier model ===

J48 pruned tree
------------------

RegionName = Andaman and Nicobar Islands: 0 (1.0)
RegionName = Andhra Pradesh: 0 (15.0)
RegionName = Assam: 0 (1.0)
RegionName = Bihar: 0 (5.0)
RegionName = Chandigarh: 0 (8.0)
RegionName = Chhattisgarh: 0 (8.0)
RegionName = Delhi: 1 (25.0)
RegionName = Goa: 0 (1.0)
RegionName = Gujarat: 0 (7.0)
RegionName = Haryana: 0 (23.0)
RegionName = Himachal Pradesh: 0 (6.0)
RegionName = Jammu and Kashmir: 0 (18.0)
RegionName = Jharkhand: 0 (1.0)
RegionName = Karnataka: 0 (18.0)
RegionName = Kerala: 1 (57.0)
RegionName = Ladakh: 0 (20.0)
RegionName = Madhya Pradesh: 0 (6.0)
RegionName = Maharashtra: 1 (18.0)
RegionName = Manipur: 0 (3.0)
RegionName = Mizoram: 0 (2.0)
RegionName = Odisha: 0 (11.0)
RegionName = Puducherry: 0 (9.0)
RegionName = Punjab: 0 (18.0)
RegionName = Rajasthan: 0 (24.0)
RegionName = Tamil Nadu: 0 (20.0)
RegionName = Telengana: 0 (25.0)
RegionName = Uttar Pradesh: 1 (23.0)
RegionName = Uttarakhand: 0 (12.0)
RegionName = West Bengal: 0 (9.0)

Number of Leaves  :     29

Size of the tree :      30
```

Potential hotspots in INDIA predicted by the model:

Predicted hotspots
Andhra Pradesh
Delhi
Kerala
Madhya Pradesh
Maharashtra
Rajasthan
Tamil Nadu
Telengana
Uttar Pradesh

In detail:

SNo	State/UnionTerritory	Predicted	Total Positive Cases	Positive cases preference		
11	Andhra Pradesh	Hotspot		1	226	18.94737
12	Andhra Pradesh	Hotspot		1	266	17.69912
13	Andhra Pradesh	Hotspot		1	305	14.66165
14	Andhra Pradesh	Hotspot		1	348	14.09836
50	Delhi	Hotspot		1	219	0
51	Delhi	Hotspot		1	445	103.1963
52	Delhi	Hotspot		1	503	13.03371
53	Delhi	Hotspot		1	523	3.976143
54	Delhi	Hotspot		1	576	10.13384
55	Delhi	Hotspot		1	576	0
56	Delhi	Hotspot		1	669	16.14583
106	Kerala	Hotspot		1	286	0
107	Kerala	Hotspot		1	295	3.146853
108	Kerala	Hotspot		1	306	3.728814
109	Kerala	Hotspot		1	314	2.614379
110	Kerala	Hotspot		1	327	4.140127
111	Kerala	Hotspot		1	336	2.752294
112	Kerala	Hotspot		1	345	2.678571
124	Madhya Pradesh	Hotspot		1	229	38.78788
125	Madhya Pradesh	Hotspot		1	229	0
126	Madhya Pradesh	Hotspot		1	229	0
127	Maharashtra	Hotspot		1	335	0
128	Maharashtra	Hotspot		1	490	46.26866
129	Maharashtra	Hotspot		1	490	0
130	Maharashtra	Hotspot		1	748	52.65306
131	Maharashtra	Hotspot		1	868	16.04278
132	Maharashtra	Hotspot		1	1018	17.28111
133	Maharashtra	Hotspot		1	1135	11.49312
170	Rajasthan	Hotspot		1	200	19.76048

172	Rajasthan	Hotspot	1	274	37
173	Rajasthan	Hotspot	1	288	5.109489
174	Rajasthan	Hotspot	1	328	13.88889
175	Rajasthan	Hotspot	1	381	16.15854
176	Tamil Nadu	Hotspot	1	309	0
177	Tamil Nadu	Hotspot	1	411	33.00971
178	Tamil Nadu	Hotspot	1	485	18.00487
179	Tamil Nadu	Hotspot	1	571	17.73196
180	Tamil Nadu	Hotspot	1	621	8.756567
181	Tamil Nadu	Hotspot	1	690	11.11111
182	Tamil Nadu	Hotspot	1	738	6.956522
185	Telengana	Hotspot	1	269	69.18239
186	Telengana	Hotspot	1	321	19.33086
187	Telengana	Hotspot	1	364	13.39564
188	Telengana	Hotspot	1	427	17.30769
189	Telengana	Hotspot	1	427	0
195	Uttar Pradesh	Hotspot	1	227	30.45977
196	Uttar Pradesh	Hotspot	1	305	34.36123
197	Uttar Pradesh	Hotspot	1	305	0
198	Uttar Pradesh	Hotspot	1	343	12.45902
199	Uttar Pradesh	Hotspot	1	361	5.247813

Conclusion

The predicted regions are to be considered to be as potential hotspots to the spread of CoViD19 pandemic. These regions are expected to might become epicentres of the pandemic in the Indian subcontinent. This prediction might help us to take early measures to control and monitor the current of the spread within and outer to the regions.

Delhi	}	Actual hotspots
Telangana	}	
Maharashatra	}	
Tamil Nadu	}	Potential hotspots
Uttar Pradesh	}	
Rajasthan	}	

Lets contribute our part by any possible means to fight against the pandemic, primarily with personal hygiene.

Stay positive. Stay home. Stay safe.

References

- Read more at: https://economictimes.indiatimes.com/news/international/world-news/italy-to-quarantine-milan-venice-and-other-regions-media/articleshow/74533581.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst
- https://www.nbcnews.com/health/health-news/italy-has-world-class-health-system-coronavirus-has-pushed-it-n1162786
- WHO report- https://www.who.int/docs/default-source/wrindia/situation-report/india-situation-report-8bc9aca340f91408b9efbedb3917565fc.pdf?sfvrsn=5e0b8a43_2
- https://www.socscistatistics.com/descriptive/histograms/
- https://ncalculators.com/statistics/skewness-calculator.htm
- http://www.alcula.com/calculators/statistics/box-plot/

Attachments:

BIN203: DATA MINING & MACHINE LEARNING

PROJECT ACTIVITIES

STATEMENT OF INDIVIDUAL LEARNING & CONTRIBUTION

Group Project Title: COVID 19- potential hotspots

Specific Learning & Contributions to the Group Project:
1. Data collection

2. Data preprocessing :

- adding and removing attributes

- log transformation

3. Building models like Naiive bayes and SMO in the weka explorer

4. Coordinating the team and assigning work amongst us after dividing it as comfortable for all.
5. Also learnt to use php and python scripts for getting the dataset to the required format, and contributed in the same.

Name:	Rating:
1. R. Shashank	Excellent
2. R. Nandhakumar	Excellent

Name: T.M.N. Yatindra Pravanan

Signature: *Yatindrapravanan*

BIN203: DATA MINING & MACHINE LEARNING

PROJECT ACTIVITIES

STATEMENT OF INDIVIDUAL LEARNING & CONTRIBUTION

Group Project Title: COVID 19- potential hotspots

Specific Learning & Contributions to the Group Project:

- Study online to figure a feasible methodology to work the discussed problem statement
- Collecting data sets
- Ensuring the data preprocessing
- Observing
    - Training data results
    - Test data results
    - Prediction results
- Collecting all other observations from team mates
- Labeling them and Compiling them
- Preparing a report on the project
- Helping the buddies anytime for anything they needed me
- Exploring and understanding the features of WEKA explorer was a very good learning in the curve.

Name:		Rating:
1.	R. Shashank	Excellent
2.	T.M.N. Yatindra Pravanan	Excellent

Name: R. NandhaKumar

Signature: *r_ndk*___

BIN203: DATA MINING & MACHINE LEARNING

PROJECT ACTIVITIES

STATEMENT OF INDIVIDUAL LEARNING & CONTRIBUTION

Group Project Title: COVID 19 - potential hotspots

Specific Learning & Contributions to the Group Project:

1. Various Data cleaning processes and factors to consider.
2. Data pre-processing, classification and clustering model building using WEKA software.
3. Analysis of classification algorithms to choose the correct algorithm for the problem statement.
4. The files were converted into the required format using Python libraries.

Name of team member	**Rating**
Yatindra Pravanan	Excellent
Nandhakumar	Excellent

Name of student: R Shashank

Signature: Shashank.