COVID-19 INDIA DATASET

IndoML-2021

JANUARY 2022

-PREPARED FOR-

Datathon-IndoML-2021

-PREPARED BY-

ANURAV MODAK

CHADARAM VIVEK

PRAKASH NAGAMANI

CONTENTS

ABSTRACT

“Since Covid-19’s emergence in India, most states were affected dreadfully. The reasons for wide spread of this infectious virus had many reasons and factors which varied from state-to-state. We used data extracted automatically from daily health bulletins published by state governments to a) breakdown the behaviour of covid waves in each state and b) analyse the reasons for surges in covid cases c) to analyse and develop real life machine learning models for prediction. We learned that major factors for surge in cases were elections, festivals, restrictions, relaxations set by the state governments and poor initial vaccination drives. But in later phases less fatality rates were achieved by huge vaccinations drives and herd immunity."

INTRODUCTION

A) PROBLEM DEFINITION:

The COVID-19 India Dataset is one of the most comprehensive datasets on the pandemic in India. It aggregates data from health bulletins published online daily by governments of major Indian states. This Datathon event challenges you to flex your brains on this dataset and come up with your own models for analysis, prediction, and insights on the evolution of the pandemic in India.

B) OBJECTIVES:

The main objectives of the analysis are listed below: -

1. To Identify Cause and Effect of different covid-waves.
    a. To identify different phases in the dataset.
    b. To discuss cause and effects in brief.
    c. Conclusion.
2. To identify the merits of vaccine and lockdown to reduce hospitalisation and fatality rate.
    a. To justify the advantages of vaccine on general terms.
    b. Merits of Lockdown.
    c. Conclusion.
3. To identify relation, association to develop machine learning models to predict future values.
    a. To identify best attributes to predict cumulative positive cases and cumulative deaths.
    b. To identify and choose best models for prediction.
    c. To identify the factors which affects cumulative cases and deaths among different states.
    d. Conclusion.

TEAM DETAILS

USERNAME: anuravmodak

REGISTERED EMAIL FOR TEAM: anuravmodak1@gmail.com

1. NAME: ANURAV MODAK
   COLLEGE EMAIL: 121910312001@gitam.in
   REGISTERED EMAIL: anuravmodak1@gmail.com
   ALTERNATIVE EMAIL: anuravmodak1@gmail.com
   COLLEGE: GITAM UNIVERSITY
   COLLEGE ID: 121910312001
   MOBILE: 7013142661

2. NAME: CHADARAM VIVEK
   COLLEGE EMAIL: 121910312046@gitam.in
   ALTERNATIVE EMAIL: chadaram79@gmail.com
   COLLEGE: GITAM UNIVERSITY
   COLLEGE ID: 121910312046
   MOBILE: 9502777255

3. NAME: PRAKASH NAGAMANI
   COLLEGE EMAIL: 121910312048@gitam.in
   ALTERNATIVE EMAIL: prakashnagamani01@gmail.com
   COLLEGE: GITAM UNIVERSITY
   COLLEGE ID: 121910312048
   MOBILE: 7659827607

STATISTICAL ANALYSIS

OBJECTIVE: To Identify Cause and Effect of different covid-waves.

Phase-1: April-2020 to January-2021

1. It was seen in some states like West Bengal where first wave lasted till December-2020 this happened when during the time when state lockdown and curfew were relaxed and full consent was given to celebrate Durga Puja festival.
2. In Delhi, there was also surge in cases during the same period but there could be two main reasons. First one being, unlocking process started which includes **people going outside** or to their **workplace, less rate of vaccination, mass gathering on festivals, revenge vacation etc**. Second could be strong protest going at the borders of Delhi where people from different states came and occupied streets, national highways and expressways which led to increase in positive cases.

   As the protest was mainly originated from Punjab, we were also able to see same trend in total positive cases in Punjab during from August 2020 to September-2020, refer graph below **Fig-1**.

To see full sized image, Click Here

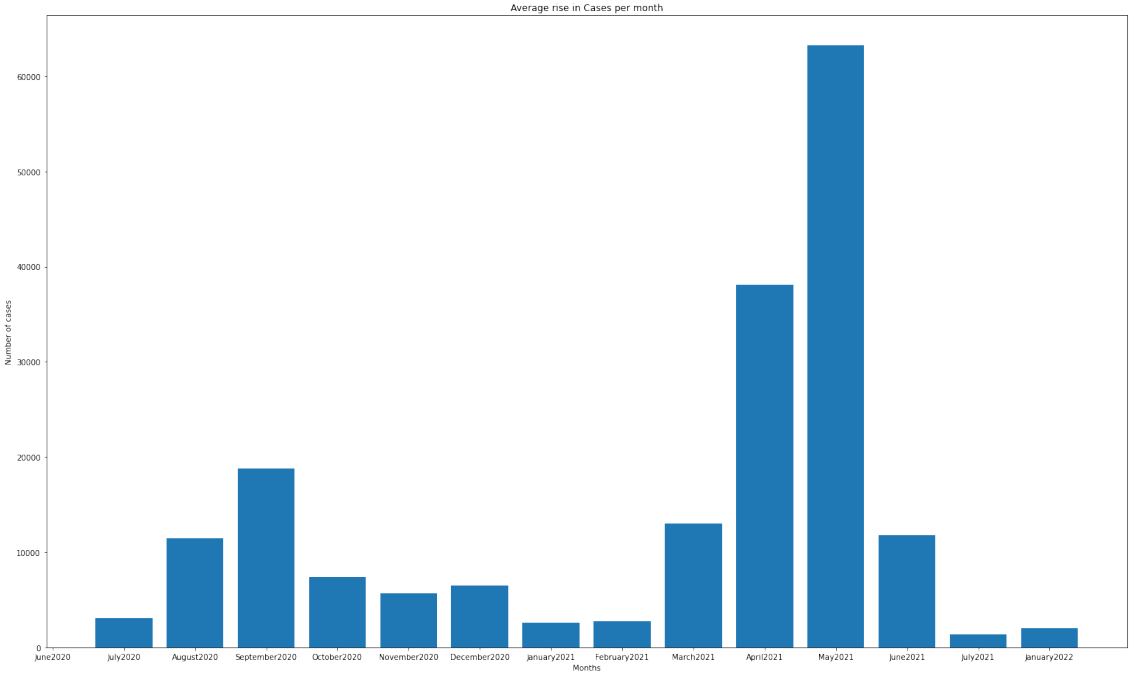

Fig-1: The graph shows "*Average rise in cases per month".*

On the same hand number of covid cases in Tamil Nadu was seen decreasing from July-2020 to December-2020. One primary reason is that the government could achieve that because the lockdown was implemented strictly till December2020.

Source: LINK

3. In Uttarakhand, during September-2020 received high hotel bookings, tourist as many neighboring states started their unlocking process as it is shown in **Fig-2**.

Phase-2: Feb-2021 to January-2022

All states except Uttarakhand saw stable atmosphere after first wave as positive cases were under control between January-2021 to March-2021, but at the same time every state witnessed sudden rise in cases from a new variant, some reasons are listed state wise:

1. During March 2021 to April-2021 we saw surge in covid cases as we can see from graph **Fig-2**. One of the major reason could be grand

celebration of Kumbh mela which is held during January 14th to April 27th . **Source:** LINK

To see full sized image, Click Here

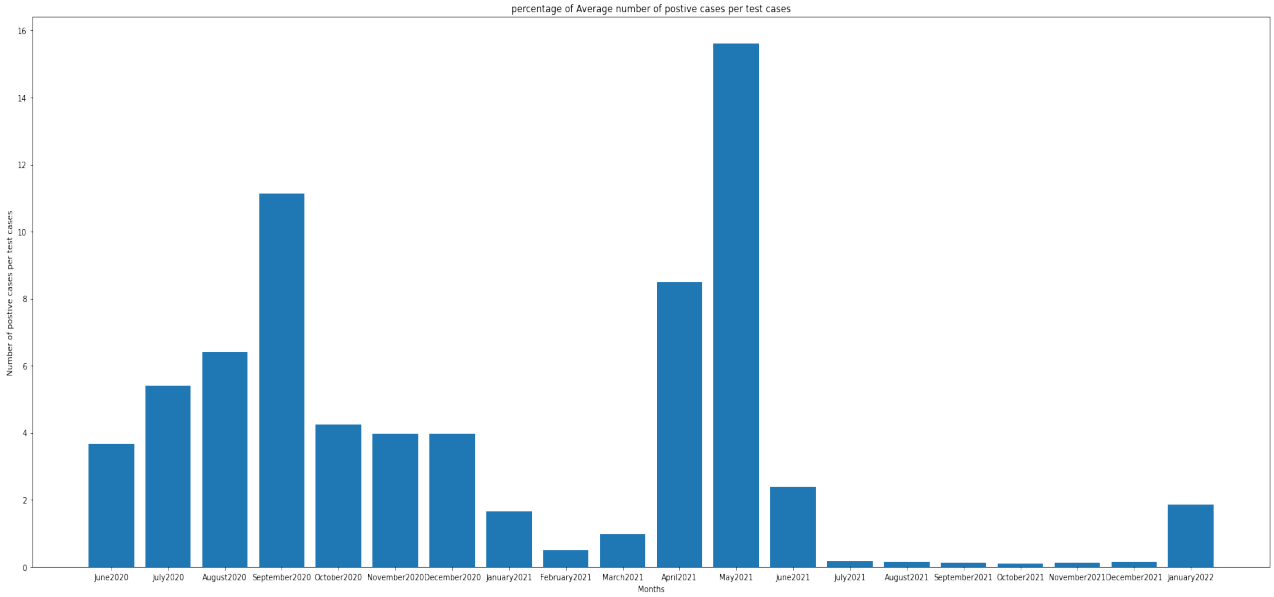

Fig-2: The graph above **shows "Average number of positive"** in Uttarakhand**.**

2. Most of the states like Telangana, Tamil Nadu, Haryana, Punjab, Goa, Kerala, Punjab, Delhi, West Bengal, Karnataka, claims festival celebration like Holi or some other festival and revenge vacation started from Feb-2021, where vacation and staycation during long weekends pushed the virus towards community spread. From the travelling data from Kerala directly shows there was a rapid increase in people travelling both domestically and internationally from January -2021 to Feb-2021.

   As the vaccination caught pace, it was seen that the rate of fatality, need hospitalisation also went down.
3. In Haryana, during April-2021 to May-2021 the **average no. people that got 1st dose per month are 1719710** and average **no. of people that got 2nd dose are 1328711**. Vaccinations started around January 2021, therefore by April approximately **51,59,132 no. of people got 1st doses** and **39,86,133 no. of people got fully vaccinated**. The total population of Haryana is 25,351,462

   i.e., only **20% got their 1st dose and 15% got fully vaccinated**.

   (**Source**: LINK)

4. In Karnataka, it was seen from **Fig-3** that there was an increase in positive cases and fatality rate from July 2021 to October 2021, political rallies were conducted which resulted in mass gathering for Legislative assembly elections which were held during a 30th October 2021 to 2nd November 2021.

To see full sized image, Click Here

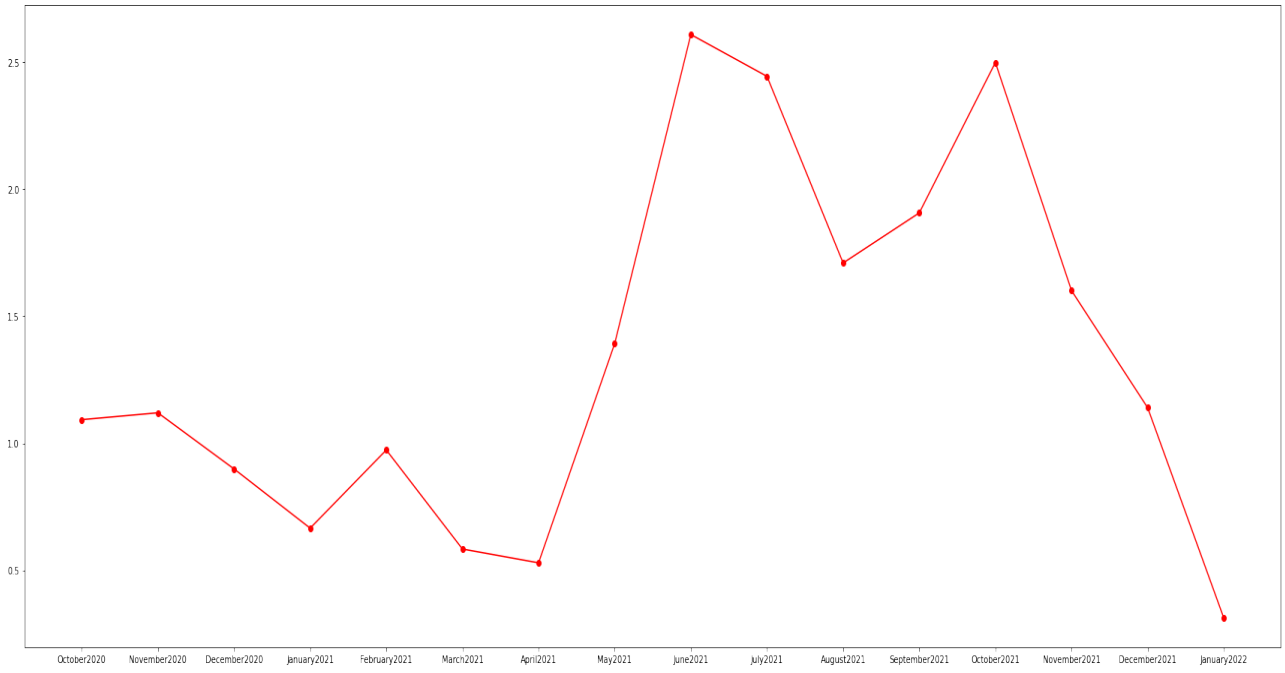

Fig-3: The graph above is shows ***Trend of Positive cases per month in Karnataka***

5. Election in West Bengal and political rallies conducted during January-2021 and Feb-2021 can be considered as super spreader of covid in the respective states. If we take our analysis as foundation as shown in **Fig-4**, we directly found that West Bengal saw an exponential rise in cases i.e., one of the highest among all states during this period. In addition to that poor vaccination rate and insufficient medical infrastructure in West Bengal were also the major reasons which compiled together bringing a strong wave of positive cases and fatalities.

To see full sized image, Click Here

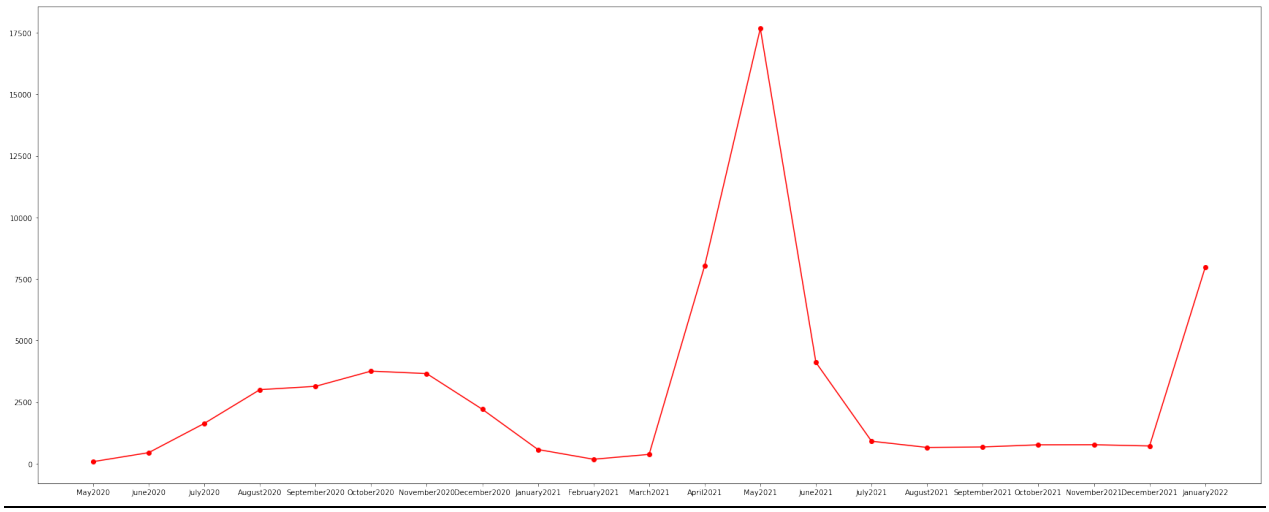

Fig-4: The graph shows *trend of new cases in each month* in West Bengal.

6. In Delhi, during the second wave it was found that many covid care facilities and covid health centres were not occupied completely even though there were shortage of bed and people preferred private hospitals. Even though in second wave more medical amenities were demanded where it seems that most govt. facilities failed to provide those demands which made people to choose private hospitals over them.

Stats of medical facilities available for covid patients in Delhi:

Month	Covid beds per million
April -2021	681
May-2021	1180
June-2021	1188

Table: 1.1.1

Stats of medical facilities available for covid patients in West Bengal:

Month	Covid beds per million	ICU beds per million	Ventilators per million
April -2021	87	18	8
May-2021	273	38	14
June-2021	223	29	14

Table: 1.1.2

From **table 1.1.1** and **table 1.1.2** clearly shows that the insufficient medical amenities in Delhi and West Bengal were likely to be the reason behind fatalities and high discomfort during pandemic among general public.

7. In Kerala, it was seen from the **Fig-5** that second wave lasted longer than all other states as the state govt. relaxed the lockdown in the state after June-2021 and grand celebration of Eid made situation worst once again for Kerala as we can see there was an increase in active cases from August-2021 to October-2021.

To see full image, Click here

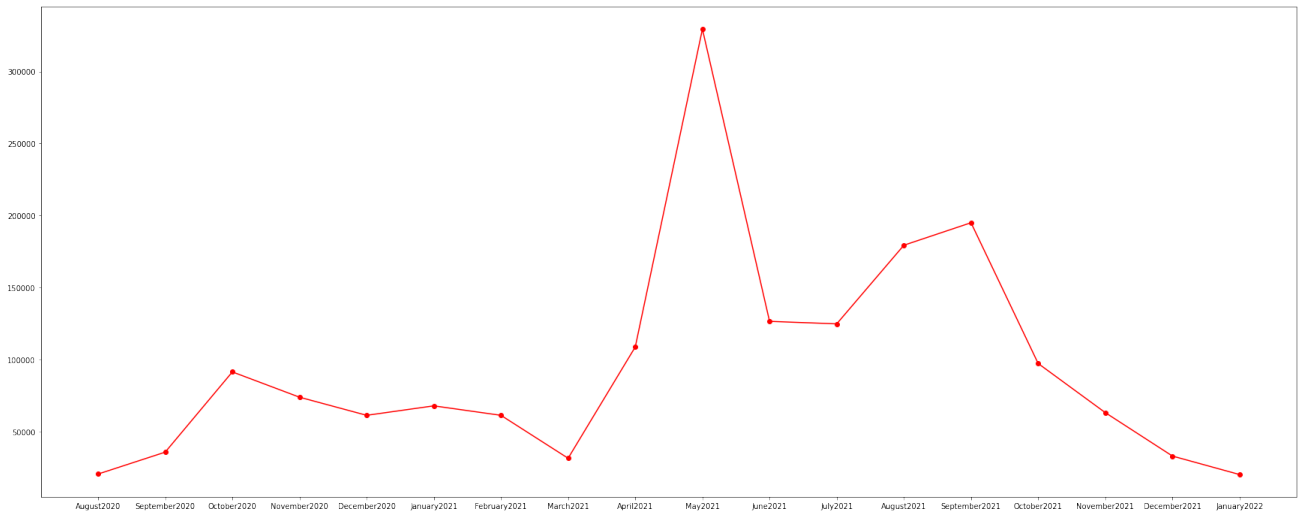

Fig-5: The graph shows average cumulative active cases on monthly basis in Kerala.

CONCLUSION:

Festivals, elections, mass gathering were some of the major reasons contributed to massive widespread virus from April-2021 to June-2021 in most states. But it was also seen that on the same reasons even some states like Kerala witnessed increase in new cases till December-2021.

Poor medical infrastructure and less vaccination added to the misery of common people.

OBJECTIVE: To identify the merits of vaccine and lockdown to reduce hospitalisation and fatality rate.

1. From the analysis it is clearly evident that most of the states with increasing vaccination witnessed downfall in number of positive cases.
2. In Punjab, it was noted that when the age group 18-44 years started to catch up high vaccination rate there was sudden downfall of positive cases from May-2021 to June-2021.

NOTE: Vaccination was not the only reason behind the sudden downfall of cases, herd immunity had major contribution towards the downfall, but vaccination helped to bridge the remaining gap left to curb the virus.

Refer the two figures:

1.**First image (Fig-6)** below is showing **daily vaccination rate** of age group 18 to 44 in Punjab.

2. **Second image (Fig-7)** image below is showing **average number of positive cases** per month in Punjab.

To see full size image, Click Here

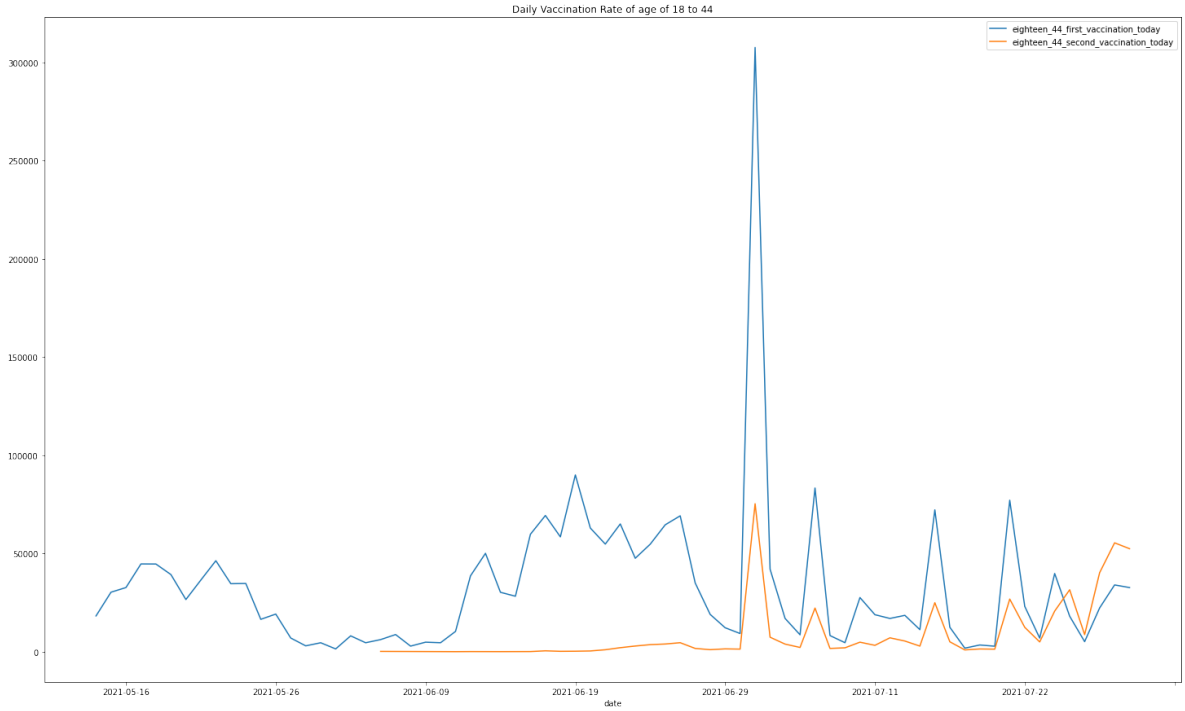

Fig-6: The graph shows Daily Vaccination Rate of age of 18 to 44.
To See full sized image, Click Here

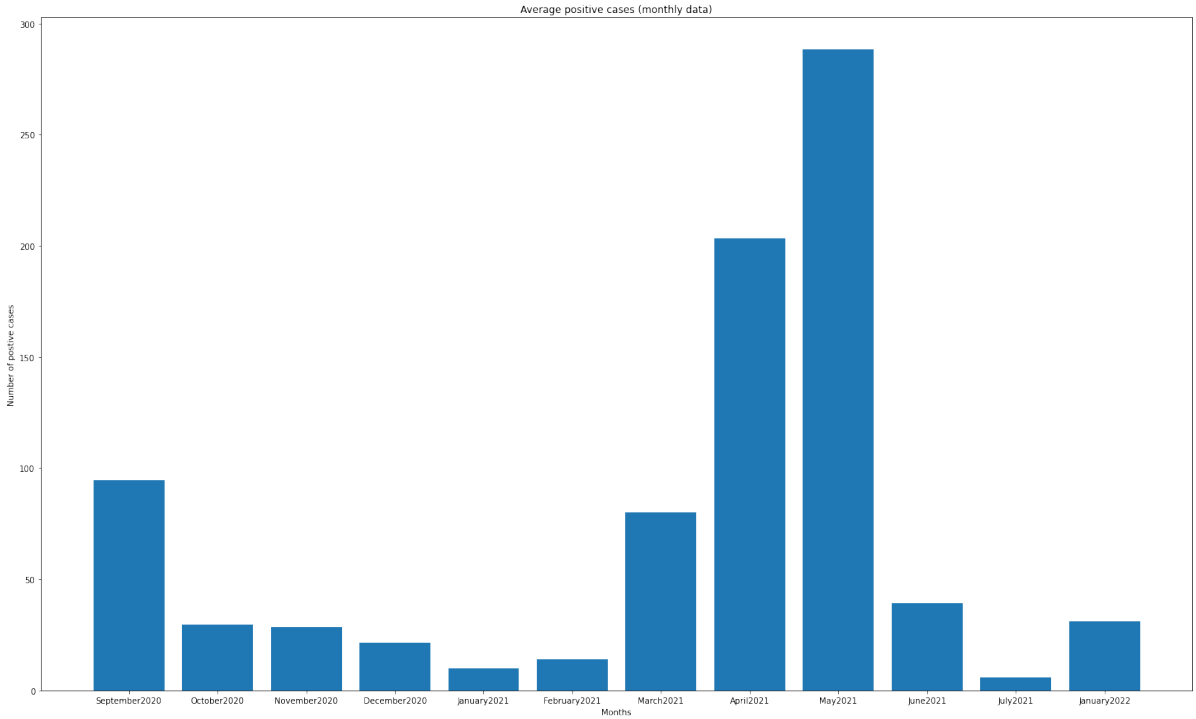

Fig-7: The graph shows **average number of positive cases** per month in Punjab.

3. In Delhi also similar trend was seen, as more number of completed their both doses of vaccines the condition in the city stabilised.

Refer the images below:

1. **First image (Fig-8)** is showing the details of **total vaccination** v/s **first dose** v/s **second dose** on monthly basis in Delhi.
2. **Second image (Fig-9)** is showing the **trend of active cases** per month in Delhi.
3. **Third image (Fig-10)** is showing **average hospital bed occupied** per month in Delhi.

To see full image-1, Click here

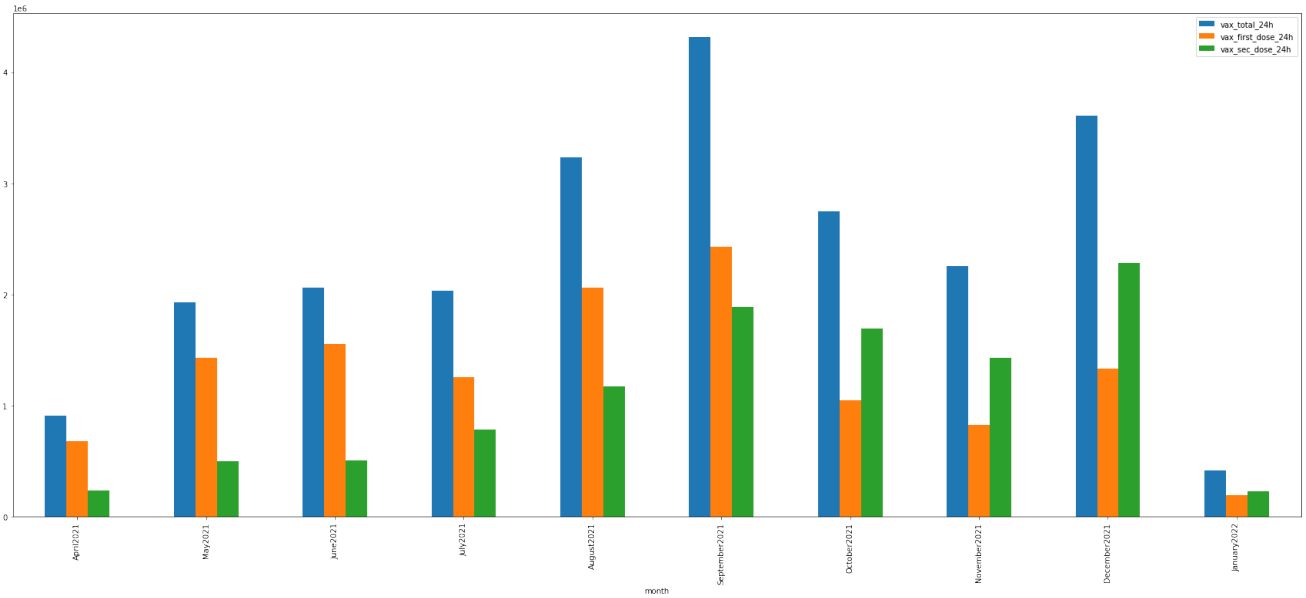

Fig-8: The graph shows the details of **total vaccination** v/s **first dose** v/s **second dose** per month in Delhi.

To see full image-2, Click Here

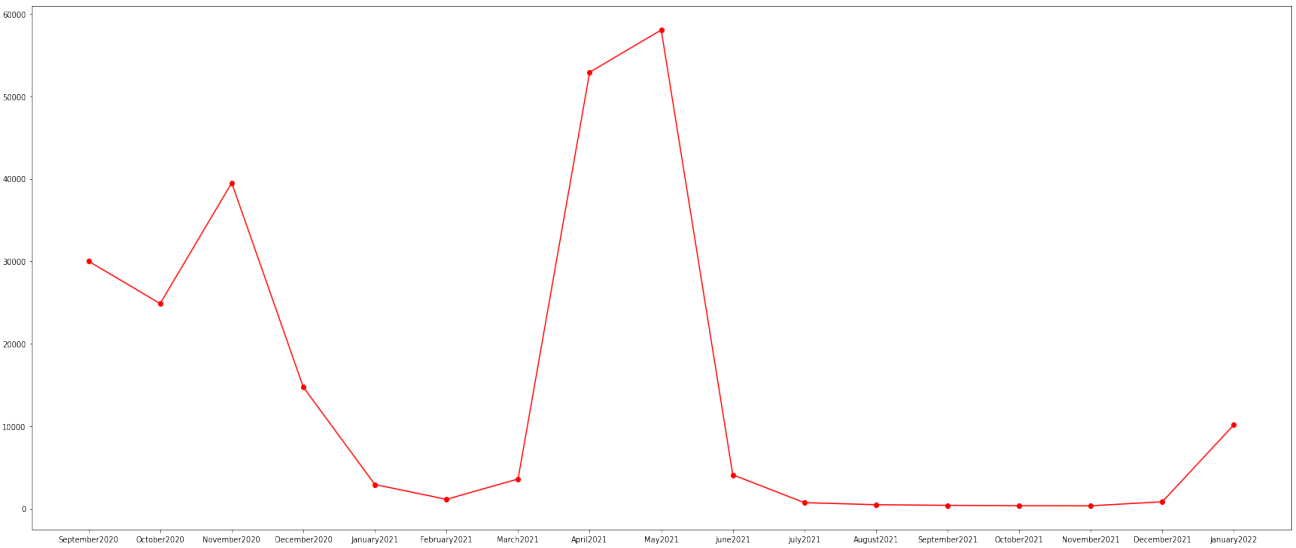

Fig-9: The graph shows the **trend of active cases** per month in Delhi.

To see full image-3, Click Here

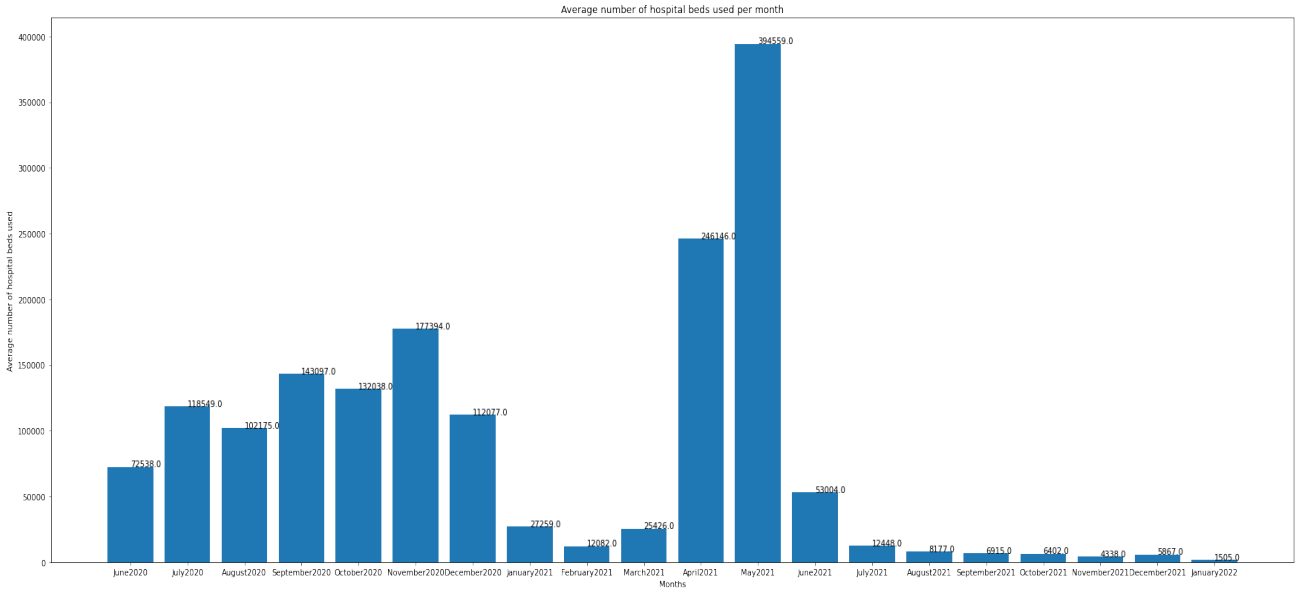

Fig-10: The graph shows average hospital bed occupied per month in Delhi.

NOTE: The analysis does not show any result on vaccine's effect on different covid variants or any upcoming variants, it is just showing the obvious advantages of vaccine on general population.

4. In Tamil Nadu number of covid cases decreased from July-2020 to December-2020 as seen in **Fig-12**. One primary reason the government could achieve that because the lockdown was implemented strictly till December-2020.

   Source: LINK:

   To see full sized image, Click Here

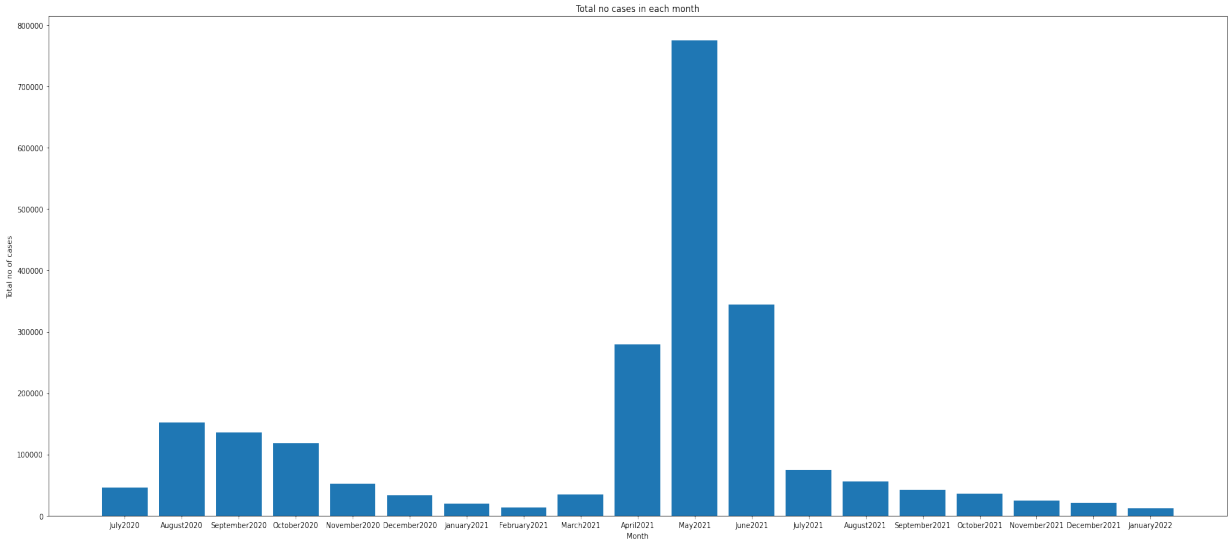

Fig-12: The graph plots **total no of positives cases** per month in Tamil Nadu**.**

5. Starting from June-2020 as shown **in Fig-13**, First wave continued till December-2020, making West Bengal to be one-of the worst hit states after first-wave in the country. This small plateau which is visible from **June-2020 to December-2020** in the above-mentioned graphs shows that it not only contributes to the number of positive case but also added to high fatality rate. West-Bengal state govt had withdrawn state-wide lockdown on **28-August-2020,** as it only contributed to more positive cases.

Source: LINK

To see full sized image, Click Here.

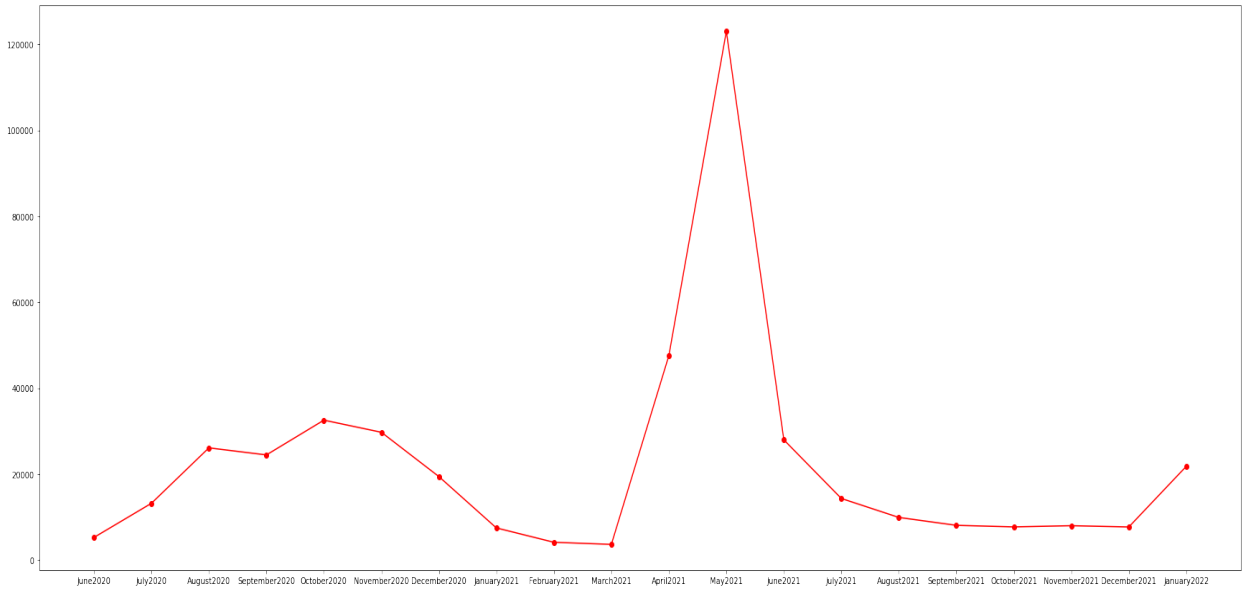

Fig-13: The graph shows **Average *of new positive cases per month*** in West Bengal

CONCLUSION:

Vaccination indeed worked as an antidote to reduce hospitalisation and thus fatality rate in many states. Strict implementation of lockdown also helped in Tamil Nadu to control the positive cases for much longer period whereas in West-Bengal as soon as lockdown was lifted, we can see rise in positive cases.

BEST FIT MODEL ANALYSIS

OBJECTIVE: To identify relation and association to develop machine learning models to predict future values.

ABSTRACT: We broadly categorised our models into two subcategories:

a) Practical Model

b) Perfect Model

Practical model shall be used to determine the factors which can be used to predict the data point in near future.

Perfect model is not meant for prediction it is developed to identify the actual factors and relations which are responsible for a particular cause within the dataset.

1) Prediction model to predict cumulative positive cases in different states:

1. Delhi: Here, we have created two types of Linear Regression model.

Practical Model:

Type of Model: Multivariate Model.

Here, our **X** and **Y** have single variable.

X=Number of days starting from 5th Jan-2022.

Y=total number of cases.

- Accuracy of model within the dataset is 85%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 23rd Jan 2022)	**1793955**
ACTUAL VALUES (As of 23rd Jan 2022)	**1728514**

Table-3.1.1

Practical Model:

Type of Model: Multivariate Model.

Here **X** has multiple variables and **Y** has single variable.
X=

i) Number of days starting from 6th Jan-2022.
ii) Cumulative positivity rate.

Y=total positive cases.

- Accuracy of model within the dataset is 90%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 23rd Jan 2022)	**1914652**
ACTUAL VALUES (As of 23rd Jan 2022)	**1728514**

Table-3.1.2

Perfect Model:

Type of Model: Multivariate Model.

Here **X** has multiple variables and **Y** has single variable.

X= i) Number of days starting from 6th Jan-2022.

ii) Cumulative positivity rate.

iii)Active cases.

iv)Cumulative Deaths.

v)Cumulative Recovered.

Y=total positive cases.

- Accuracy of model within the dataset is nearly 100%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 23rd Jan 2022)	**1782496**
ACTUAL VALUES (As of 23rd Jan 2022)	**1728514**

Table- 3.1.3

As, we can see from **table:3.1.3** that perfect model is able to predict the future values with very high accuracy but we also need prior data in many fields which makes it less practical. The main purpose of developing such model is to show that prediction of cumulative cases depends upon many factors.

2. Punjab:

We combined Linear Regression model with "PolynomialFeatures" in python (read about it: Here) to fine tune our model.

Practical Model:

Type of Model: Univariate Model.

Here, X is transformed to polynomial feature with degree=2.

Here our x and y both have one variable.

X =Number of days starting from 6th Jan-2022.
Y=total positive cases

- Accuracy of the model is nearly 93% within the dataset.

Accuracy of the model outside the dataset is given below: -

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**812786**
ACTUAL VALUES (As of 27th Jan 2022)	**728042**

Table- 3.2.1

Perfect Model:

Here x has multiple variables and y has single variable.
X=

i) Number of days starting from 6th Jan-2022.
ii) Active Cases on that day.
iii) Total recovered case until the given date.

Y=total positive cases.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 25th Jan 2022)	**729480**
ACTUAL VALUES (As of 25th Jan 2022)	**728042**

Table- 3.2.2

From **table-3.2.2** it is noted that perfect model is an ideal case model which more accurate but it is also noted that we need prior data in many fields before we can develop a model, thus makes it less practical.

3) West Bengal:

Practical Model:

Type of Model: Univariate Model.

Here our **X** and **Y** both have one variable.
X =total number of days starting from 6th Jan-2022.
Y=total positive cases

- Accuracy of the model is nearly 93% within the dataset.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 25th Jan 2022)	**1918059**
ACTUAL VALUES (As of 25th Jan 2022)	**1974285**

Table- 3.3.1

Perfect Model:

Type of Model: Multivariate Model.

Here x has multiple variables and y has single variable.

X=

i) Number of days starting from 6th Jan-2022.
ii) Active Cases on that day.
iii) Total recovered case until the given date.

Y=total positive cases

- Accuracy of the model within the dataset is nearly 100%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 25th Jan 2022)	**1976941**
ACTUAL VALUES (As of 25th Jan 2022)	**1974285**

Table- 3.3.2

From **table-3.3.2** it is noted that perfect model is an ideal case model which more accurate but it is also noted that we need prior data in many fields before we can develop a model, thus makes it less practical.

4. Telangana:

Practical Model:

Type of Model: Multivariate Model.

Here, X is transformed to polynomial feature with degree=2.

X=Number of days from starting from 6th Jan-2022.
Y=Total cases.

- Accuracy of the model within the dataset is 90.25%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**777495**
ACTUAL VALUES (As of 27th Jan 2022)	**750199**

Table- 3.4.1

Perfect Model:

Type of Model: Multivariate Model.

Here, we used multivariable Linear regression.

X= i) Number of days starting from 4th Jan-2022.

ii) Total recovered cases.

iii) new cases on that day.

Y= Total cases.

- Accuracy of the model within the dataset is nearly 100%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**751383**
ACTUAL VALUES (As of 27th Jan 2022)	**750199**

Table- 3.4.2

2) Prediction model to predict cumulative death cases in different states:

1. **Delhi:** Here we have developed a Linear Regression model using multiple variables.

Practical Model:

Type of model: Univariate Model

Here our **X** and **Y** both have one variable.
X =Number of days starting from 5th Jan-2022.
Y=Cumulative death cases.

- Accuracy of the model is 86.24%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 23rd Jan 2022)	**31009**
ACTUAL VALUES (As of 23rd Jan 2022)	**25586**

Table-4.1.1

Perfect Model:

Type of model: Multivariate Model

Here **X** has multiple variables and **Y** has single variable.
X=

i) Number of days starting from 6th Jan-2022.
ii) Total recovered case until the given date.
iii) Number of active cases.

Y=total death cases.

NOTE: From correlation coefficient we found out that among all the variables in X, "total death cases" has strongest relation with "Total recovered case".

- Accuracy of the model is nearly 100%

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 23rd Jan 2022)	**28626**
ACTUAL VALUES (As of 23rd Jan 2022)	**25586**

Table-4.1.2

From **table-4.1.2** it is noted that perfect model is an ideal case model which more accurate but it is also noted that we need prior data in many fields before we can develop a model, thus makes it less practical

2. Punjab: We combined Linear Regression model with "PolynomialFeatures" in python (read about it: Here) to fine tune our model.

Practical Model:

Type of Model: Univariate Model

Here, X is transformed to polynomial feature with degree=1.

Here our **X** and **Y** both have one variable.
X =Number of days starting from 6th Jan-2022.
Y=total positive cases

- Accuracy of the model is nearly 93% within the dataset.

Accuracy of the model outside the dataset is given below: -

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**16188**
ACTUAL VALUES (As of 27th Jan 2022)	**17081**

Table- 4.2.1

Practical Model:

Type of Model: Multivariate Model

Here, X is transformed to polynomial feature with degree=1.

Here our **X** and **Y** both have one variable.

X =Number of days starting from 6th Jan-2022.
Y=total positive cases

- Accuracy of the model is nearly 84% within the dataset.

Accuracy of the model outside the dataset is given below: -

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**16050**
ACTUAL VALUES (As of 27th Jan 2022)	**17081**

Table- 4.2.1

Perfect Model:

Type of Model: Multivariate Model

Here **X** has multiple variables and **Y** has single variable.

X=

i) Number of days starting from 6th Jan-2022.

ii) Active Cases on that day.

iii)Cumulative cases till that day.

iv) Total recovered case until the given date.

Y=total positive cases.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**17081**
ACTUAL VALUES (As of 27th Jan 2022)	**17081**

Table- 4.3.1

From **table-4.3.1** it is noted that perfect model is an ideal case model which more accurate but it is also noted that we need prior data in many fields before we can develop a model, thus makes it less practical**.**

3) West Bengal:

Here we have conducted test on both Single valued attributes and Multi valued attributes both of them gave nearly same result.

Practical Model:

Type of Model: Univariate Model

Here our x and y both have one variable.
X =total number of days starting from 6th Jan-2022.
Y=total death cases

- Accuracy of the model is nearly 93% within the dataset.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 25th Jan 2022)	**23026**
ACTUAL VALUES (As of 25th Jan 2022)	**20411**

Table-4.3.1

Practical Model:

Type of Model: Multivariate Model

Here **X** has multiple variables and **Y** has single variable.

X=

i) Number of days starting from 6th Jan-2022.
ii) Active Cases on that day.
iii) Total cases till the given date.

Y=total death cases.

- Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 25th Jan 2022)	**23367**
ACTUAL VALUES (As of 25th Jan 2022)	**20411**

Table-4.3.2

4) Telangana:

Practical Model:

Type of Model: Multivariate Model

Here, **X** has multiple variables and **Y** has single variable value.

X= i) Number of days starting from 6th Jan-2022.

ii) Total cases.

Y= Total deaths.

- Accuracy of the model within the dataset is above 95%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**4587**
ACTUAL VALUES (As of 27th Jan 2022)	**4081**

Table- 4.4.1

Perfect Model:

Type of Model: Multivariate Model

Here, **X** have multiple variables and **Y** have single variable value.

X= i) Number of days starting from 6th Jan-2022.

ii)Total Recovered cases.

iii) Total cases.

iv) New cases on that day.

Y= Total deaths.

- Accuracy of the model within the dataset is nearly 100%.

Accuracy of the model outside the dataset is given below:

PREDICTION OF THE ABOVE MODEL V/S ACTUAL VALUES:

PREDICTED VALUES (As of 27th Jan 2022)	**4347**

ACTUAL VALUES (As of 27th Jan 2022)	**4081**

Table- 4.4.2

From **table-4.4.2** it is noted that perfect model is an ideal case model which more accurate but it is also noted that we need prior data in many fields before we can develop a model, thus makes it less practical.

CONCLUSION:

Broadly we have categorised our prediction models into two categories:

1) Practical Models:

These models are practical in nature and can be used in short term to predict values.

These models are developed to identify the attributes which in real life can be used to predict future trends in "cumulative cases" and "cumulative deaths"

This model contains both univariate and multivariate models.

From the above analysis our conclusion on attributes which is most often used are: -

a) To predict cumulative cases are:

"**Number of days**" and "**positivity rate**" (if available).

b) To predict cumulative deaths are:

"**Number of days**" and "**Cumulative cases**" (if available).

2) Perfect Model:

These are Ideal models that are developed intentionally to give 100% accuracy.

This helped us to truly understand the actual trend and to identify the strongest relationship to between two variables.

The attribute(s) which has the strongest association and relation to accurately predict:

1) Cumulative cases:

"**Number of days**" and "**Cumulative recovered case**".

2) Cumulative deaths: "**Cumulative recovered case".**

"**Cumulative deaths**" and "**Cumulative recovered case**" also has highest Pearson-Correlation coefficient in all available states.

Hence, it was found that "**Cumulative recovered**" was more effective in making prediction model to predict trend of fatalities and "**number of days**" and "**cumulative recovered case**" were the most effective attributes in making prediction for **"Cumulative case"**.

EXPLORATORY DATA ANALYSIS

Objective-1: To show direct relationship between population and spread of virus:

- In Kerala we found that: the analysis is directly proportional to the population of each district, the districts with *High-Positivity Rate* are the top 4 districts in terms of population in Kerala, which includes **Ernakulam**, **Kozhikode**, **Thiruvananthapuram**, **Thrissur** as shown in **Fig-14**.

- For more detailed info about Kerala please refer Jupyter notebook.

To see full sized image**,** **Click Here**

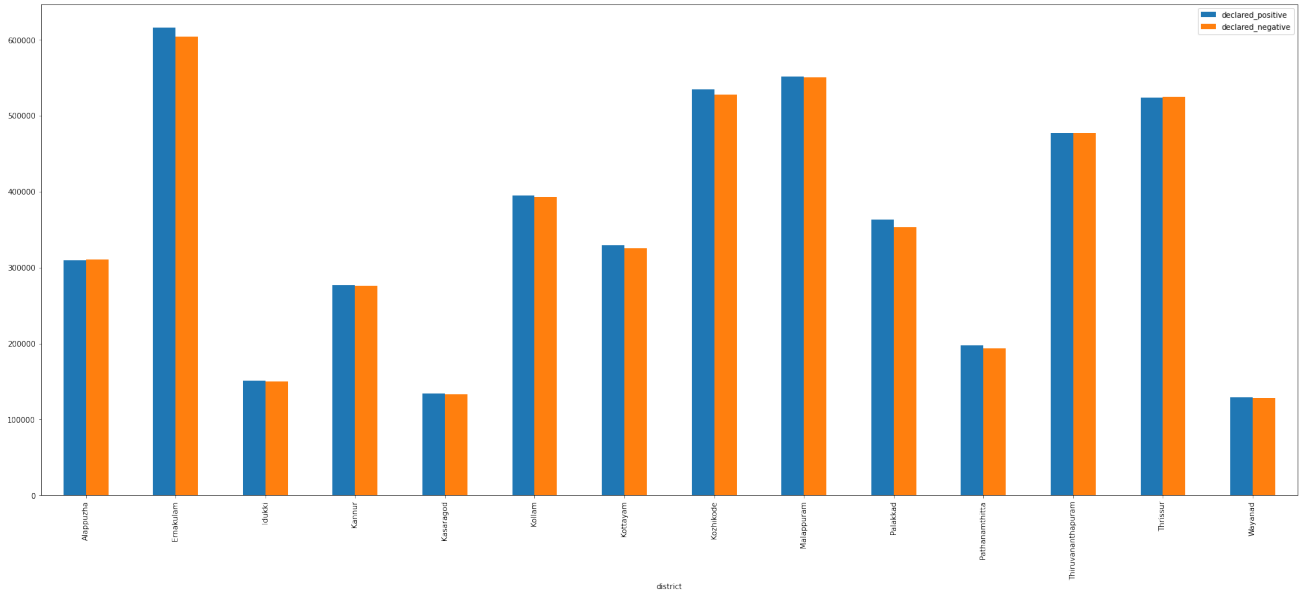

Fig-14: The graph **shows total cases positive V/S total cases negative (district wise)** in Kerala**.**

- In Tamil Nadu we saw similar trend: Chennai accounts 37.68% of total cases of the state. Then follows Coimbatore with 18.55%, Chengalpattu with 12.31%, Tiruvallur with 8.62% etc. We can observe that all these districts are mostly urban areas.

To see full sized image, **Click Here**

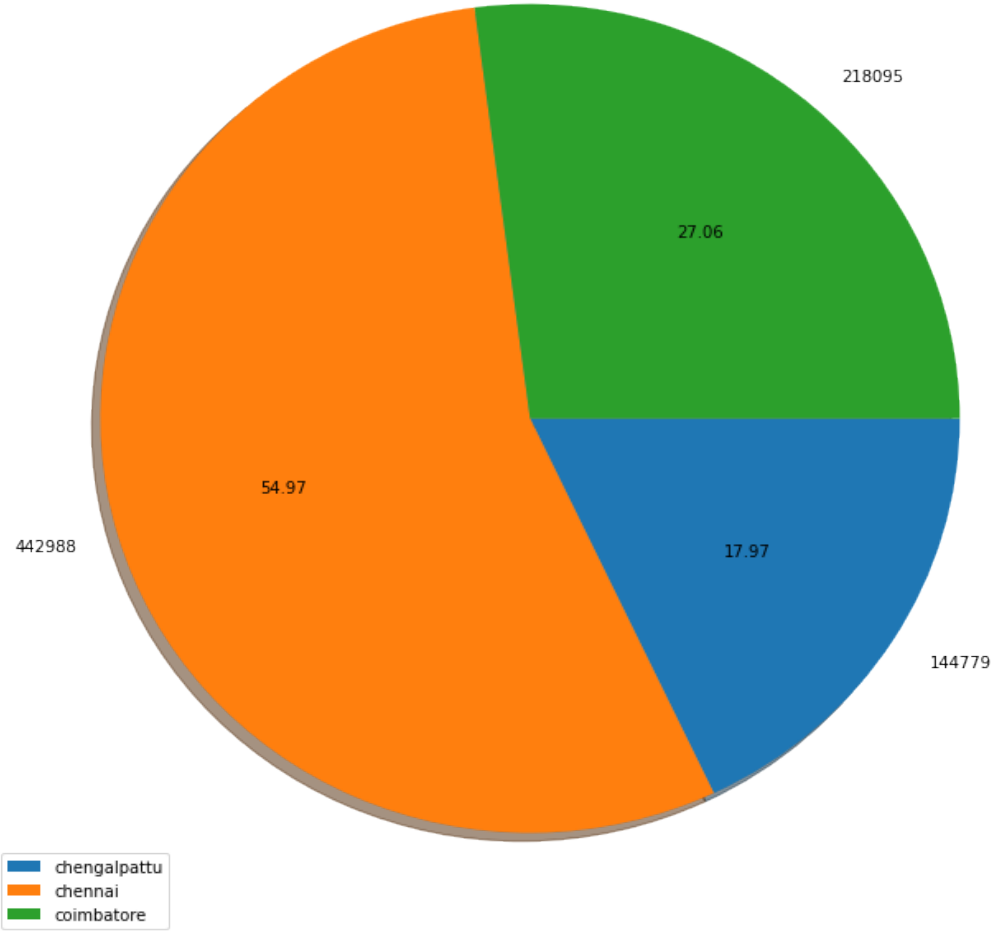

Fig-15: The pie chart **shows major cities contribution to total cases** in Tamil Nadu.

- In Punjab, 31.76% of the deaths are from major cities like Amritsar, Jalandhar, Ludhiana. The major reasons are Mass gatherings during the farmers protest and lack of covid protocols.

Objective-2: Trend in Fatalities and different comorbidities involved:

- In Tamil Nadu it was found that death rate is very high in patients of age 40 above compared to deaths in patients below 40 age group. Again, in patients of age 40 above huge number of deaths can be seen in patients with comorbidities than without patients without comorbidities.
  (**Check table 5.2.1**)

  Table 5.2.1 shows **no. of deaths in 40+ age group with vs without comorbidities.**

Fatalities category	**Number of deaths**
With comorbidity	Nearly 8000
Without comorbidity	Nearly 19000

Table 5.2.1

- **To view bar graph** Click Here

To see full sized image, Click Here

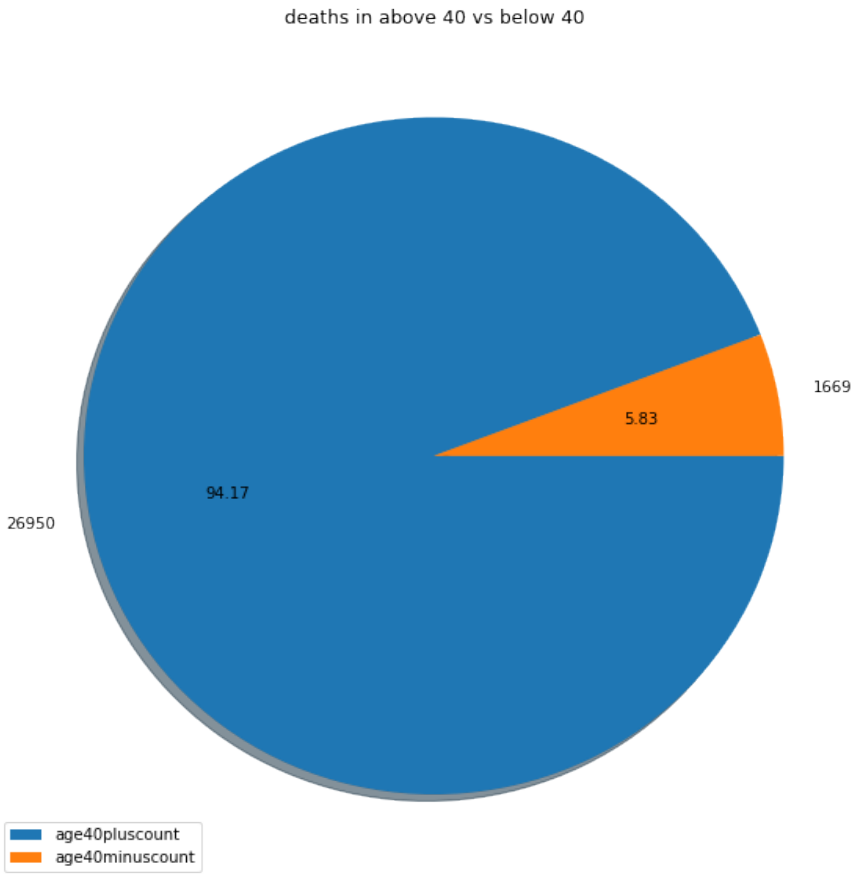

Fig-16: The pie chart depicts **the no. of deaths in 40+ age group vs 40+ age group (with comorbidity)** in Tamil Nadu**.**

- In Telangana: Unlike in the first wave, spread of virus is more in younger population. There can be multiple reasons but major reasons can be the severity of the current strain and low vaccination rate among younger population**:**
  Source: LINK

- In Karnataka: From **Fig-17** nearly 71% of deaths cases are having of Severe Acute Respiratory Infections (SARI) and

Nearly 25% of deaths are cases are having of influenza like illness (ILI).

To see full sized image, Click Here

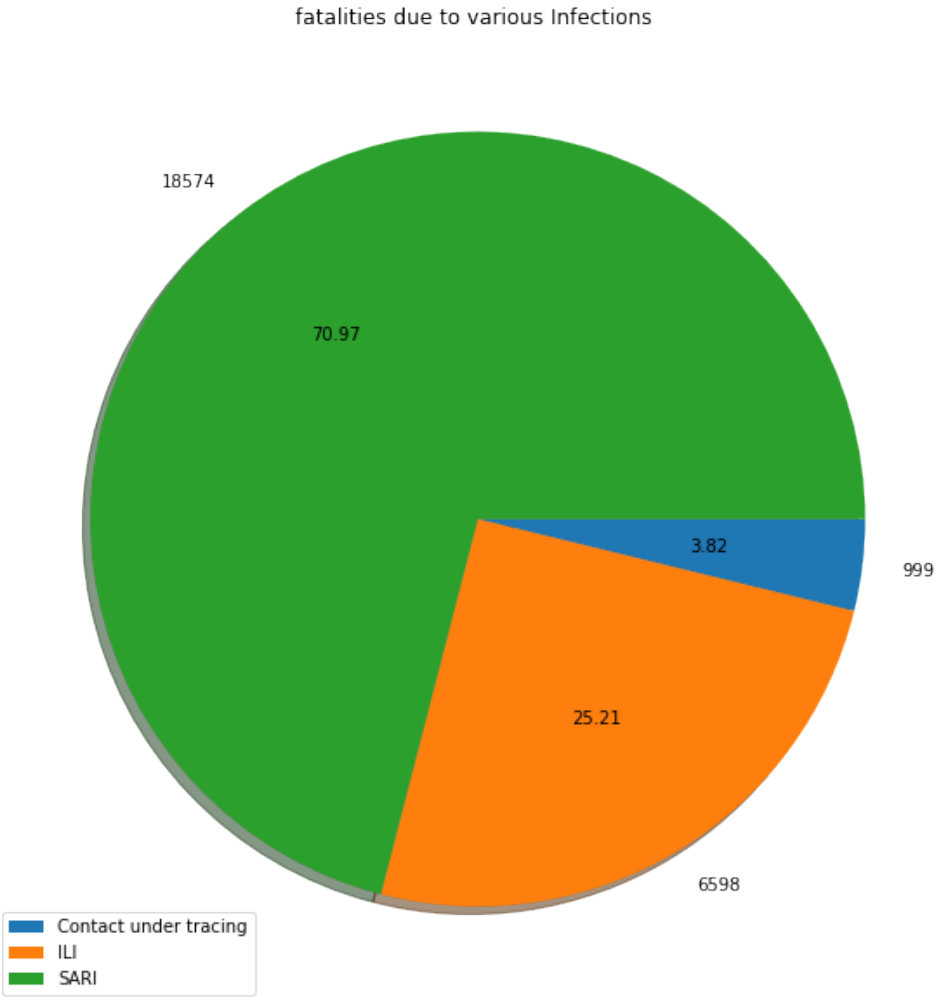

Fig-17: The pie-chart shows **Fatalities due to various infections** in Karnataka**.**

- Death rate is very high in patients of age 45 above compared to the death rates in patients below 45. Nearly 85% of deaths are of patients whose age>45. Nearly 13% of deaths are of age between 18 to 44.

To see full sized image, Click Here

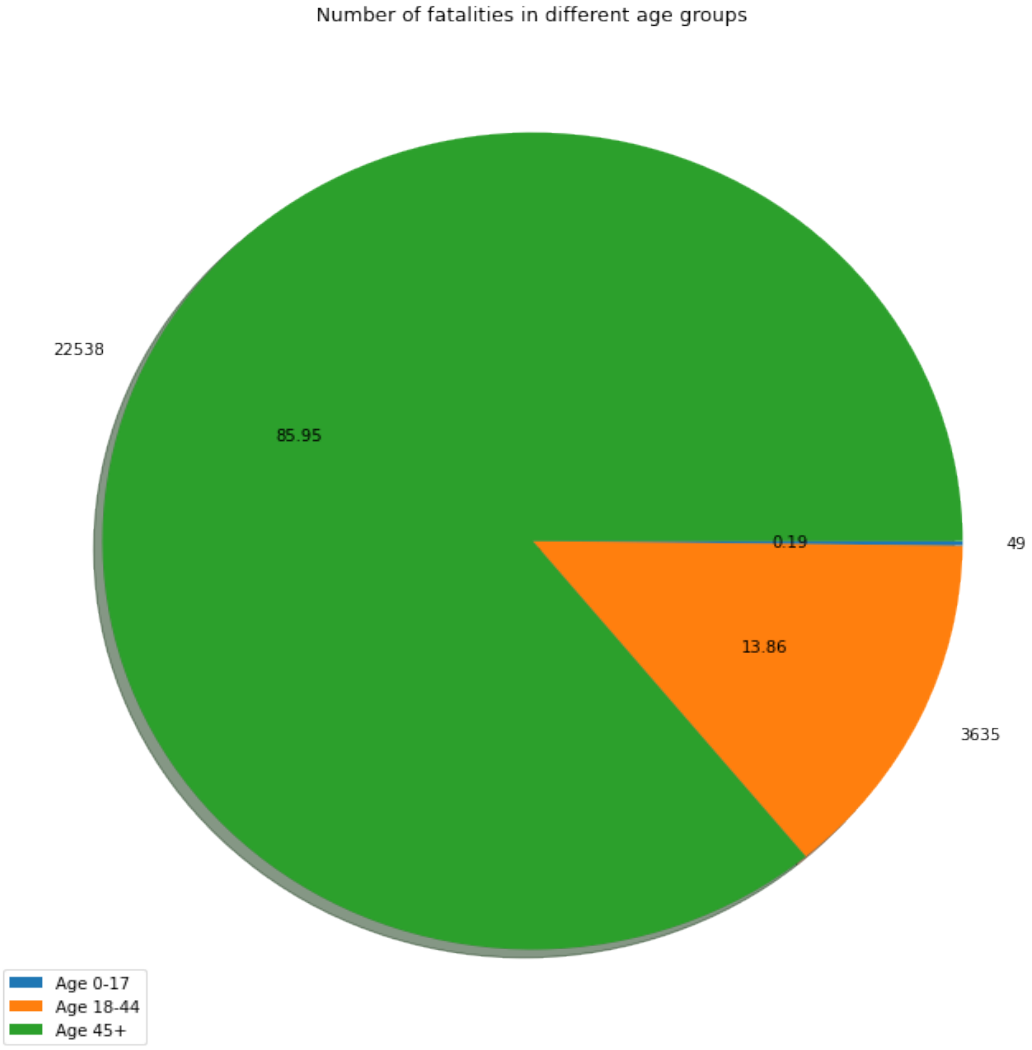

Fig-18: The pie chart shows **number of fatalities classified by age groups** in Karnataka.

Comparison in fatalities with comorbidity and without comorbidity:

- To see full sized image, Click Here

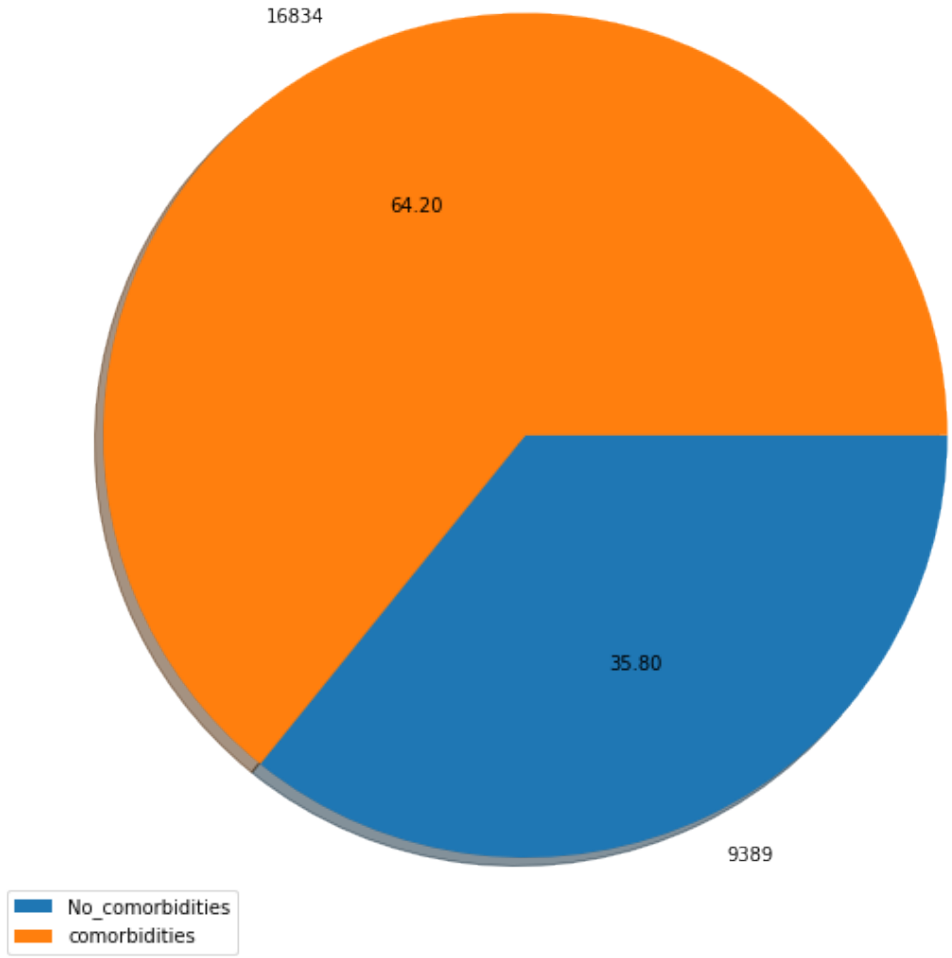

Fig-19: The graph is showing number of fatalities **with comorbidity and without comorbidity** in Karnataka.

FINAL CONCLUSION OF THE REPORT

1. Festivals, mass gathering in election rallies, poor rate of vaccination, not following strict lockdown in states were some of the major reasons for high positive case, hospitalisation and fatality.
2. Poor medical infrastructure in the country has caused a lot of discomfort to people and to the medical fraternity in this pandemic.
3. Vaccination played a major role to reduce hospitalisation and fatality.
4. It was identified that all perfect models truly displayed the actual factors behind “total cases” and “total deaths”. At the same time practical model gave us an idea on how to develop machine learning model which are feasible in real world.
5. We also identified that cities and district with big population were major contributors in positive cases and fatalities in their respective states.
6. With available data in some states, it was found that fatalities with comorbidities were more than fatalities without comorbidities.
7. With available data in some states, fatality was more in the age group 45 and above.