ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

FIGHTING EPIDEMICS AND PANDEMICS WITH DATA

Shllok Tahiliani
Trinity International School
pes.trinityis@gmail.com

Abstract

The world was completely unprepared when the COVID-19 pandemic struck, which is why the world was in chaos. We certainly hadn't learnt anything from the previous epidemics or pandemics, but now is the time to change that - to stack up our data science ammunition, and be prepared. This research paper aims to suggest a methodology for such times using the techniques - clustering and classification. The experiment was uniquely conducted to test the mentioned methodologywith a well-rounded data set, which proved fruitful. In conclusion, 'Trial and Error' is the only process that works in handling epidemics or pandemics.

Keywords: *Clustering, Classification, Epidemics, Pandemics*

INTRODUCTION

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users can translate into tangible business value [1].

Currently, data science is already being used to tackle epidemics and pandemics, such as the COVID-19 Pandemic. One example would be the apps that have been developed to monitor the spread of the disease. These generate ID on your phone and save the ID of the people you have been in contact with. If anyone tests positive, the app reviews the contact data history generated in the last few days and alerts you, to stop the spread of the disease. This way, the app predicts the spread of the disease and produces statistics of the same. There are many more such examples,like telemedicine, which are used to control to spread of the virus [2].

The problem is, data science mostly being used to control the spread of the disease, not to help the patients and/or healthcare system. The objective of this paper is to suggest data science techniques that can be used to guide the patient once they are tested positive and thus reducing the burden on the healthcare system.

Techniques:

These are the suggested techniques to be used to guide patients.

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

Clustering

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is a collection of objects based on similarity and dissimilarity between them [3].

The following points show why clustering is required in data science −

1. Scalability − We need highly scalable clustering algorithms to deal with large databases.
2. Ability to deal with different kinds of attributes − Algorithms should be capable to be applied to any kind of data such as interval-based (numerical) data, categorical, and binary data.
3. Discovery of clusters with attribute shape − The clustering algorithm should be capable of detecting clusters of arbitrary shapes. They should not be bounded to only distance measures that tend to find a spherical cluster of small sizes.
4. High dimensionality − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
5. Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
6. Interpretability − The clustering results should be interpretable, comprehensible, and usable [4].

In an epidemic/pandemic scenario, clustering will be of great use as it will help group and organize the data that is available and identify the trends, which then can be used in the next technique i.e., Classification.

Classification

Classification analysis is a data analysis task within datascience, that identifies and assigns categories to a collection of data to allow for more accurate analysis. It can be used to question, make a decision, or predict behavior through the use of an algorithm. It works by developing a set of training data that contains a certain set of attributes as well as the likely outcome. The job of the classification algorithm is to discover how that set of attributes reaches its conclusion [5].

Classification cuts down the data set into manageable groups with similarities and assists the decision-making process. Classification is of particular importance in the healthcare industry as it makes essential data easy to find and retrieve by tagging the classified groups with commonalities which in turn assist the decision-making algorithm.

Classification helps predict the outcome of a patient in an epidemic/pandemic scenario, using the data that was clustered, which will be of great help, which will help patients make decisionsand reduce the burden on doctors.

Methodology:

This is the suggested methodology to be followed when using data science to fight micro pandemic-causing enemies.

1. Data collection of previous records
   The first and the most important step in data science is to collect as much data as possible. Data forms the foundation of all the other steps. That's why adequate data must be available. An epidemic or pandemic is declared after the disease has spread to most

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

parts of the world. So as soon as it is declared, sufficient data to work on is already available. But equally important is to make sure that the data available is accurate and relevant. For example, data collected from hospitals and/or the government will be more accurate than data collected directly from patients, as most of the times patients fail to remember the exact details, while the former have written records of all the data. John Turkey, an American statistician, once said, "The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." [6] That's why data must be relevant.

2. Clustering
   After data collection, the next step is clustering. For that, the data needs to be represented in chart form (preferably, scatter chart) so that clusters can be seen clearly and marked. But for that, a lot of data is needed, which needs to be made sure of in the previous step. After that's done, just by a look at the chart, trends can be identified – the severity of the disease increases with age, most of the patients above the age of 60 were hospitalized and so on.Identifying these trends is important for the next step.
3. Creating an algorithm
   Using the trends identified, an algorithm should be created, which takes basic details of the patient, such as, age, health condition, and will guide them on what medicine to take and whether or not they need to get themselves admitted. After a few tests, this algorithm can then be expanded to an application, linked to the dataset, which itself identifies trends and suggests patients. This application can be launched to the general public, widening the scope of the project and tackling epidemics and pandemics better. Please note, the scope of this algorithm/application is not to replace doctors and/or medical professionals, but to help them by reducing the burden on them and to prevent the healthcare system from collapsing.
4. Classification
   After the algorithm/application is ready and tested, it should collect basic details about the patient and classify them into the clusters made, and, in the process, guide them. This step has to be repeated for every patient multiple times. These suggestions from the algorithm/application will certainly help reduce the burden on doctors, and they can just check the suggestions and modify them if needed.
5. Repetition
   All of these steps will have to be repeated multiple times during an epidemic or pandemic. This needs to be done as the pathogen causing the disease is mostly a new one, so scientists are still finding information about it. Various treatments and vaccines are invented, which call for the repetition of these steps. Also, it is possible the pathogen mutates, changing everything, right from symptoms to treatments.

Experiment:

To check the credibility of the above techniques and methodology, an experiment was carried out. To check how well the methodology fares in an epidemic/pandemic scenario, the experiment was based on the ongoing COVID-19 pandemic. The following is a timeline:

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

MISA Global Partnership
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

Table 1 Experiment Timeline

Event	Period
Deciding data be gathered	9th to 11th May 2021
Creating a form	15th May 2021
Distributing the form & collecting data	16th to 20th May 2021
Extracting data & clustering	21st to 23rd May 2021
Creating an algorithm	26th and 27th May 2021
Testing & Classification	28th May 2021

1. Deciding data be gathered
   First of all, it was to be decided what data needs to be collected. The following was decided to be collected:
   a. Age – To check if COVID-19 affects people of different ages differently
   b. Gender – To check if COVID-19 affects people of different genders differently
   c. The density of an area – To find out how many people in the same area tested positive before the patient
   d. Availability of a toilet – To check if the patient has more than one toilet available at their house (decides between quarantine center and home quarantine)
   e. Precautionary measures – To find out how well these measures prevent COVID-19 infection
   f. Number of people in a family – Potential spreaders or catchers
   g. Number of people met daily – Potential spreaders or catchers
   h. Current health condition – To check if it affects the severity of the infection
   i. Family history – To check if it affects the severity of the infection
   j. Diet – To check if it affects the severity of the infection
   k. Drug intake (including alcohol & tobacco) – To check if it affects the severity of the infection
   l. Whether they exercise – To check if it affects the severity of the infection
   m. Occupation – To check if it affects the severity of the infection
   n. Mode of transport – To check if it affects the severity of the infection
   o. Travel – To check if it affects the severity of the infection
   p. COVID-19 variant/mutation – To check if it affects the severity of the infection
   q. Number of times infected earlier by COVID-19 – To check if it affects the severity of the infection
   r. Season – To check if it affects the severity of the infection
   s. Festivals/Celebrations/Events – To check if it affects the severity of the infection
   t. Reports of COVID-19 related tests (HRCT, CRP, CBC etc.) – Measure of severity
   u. COVID-19 vaccination – To check if it affects the severity of the infection
   v. Other vaccination – To check if it affects the severity of the infection
   w. COVID-19 treatment/medication – A result
   x. Healthcare in that area – To check if it affects the severity of the infection
   y. Number of days infected – Measure of severity
   z. Final Outcome – A result

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

MISA Global Partnership
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

It is also possible to collect more data, such as doses of the medicine, to expand this experiment.

2. Creating a form

After the data to be collected was finalized, an online form was made to simplify the process of collecting the data. JotForm, an online form builder, was usedto create the form. The form contained 33 questions and a welcome and thank you page.

The link for viewing and/or filling the form is https://form.jotform.com/211304594729054.

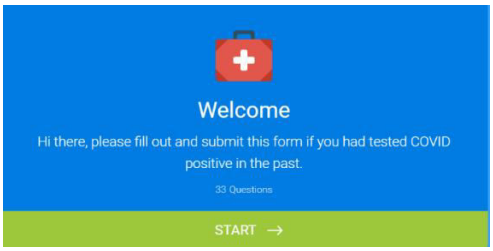


Fig. 1 Welcome Page of the form

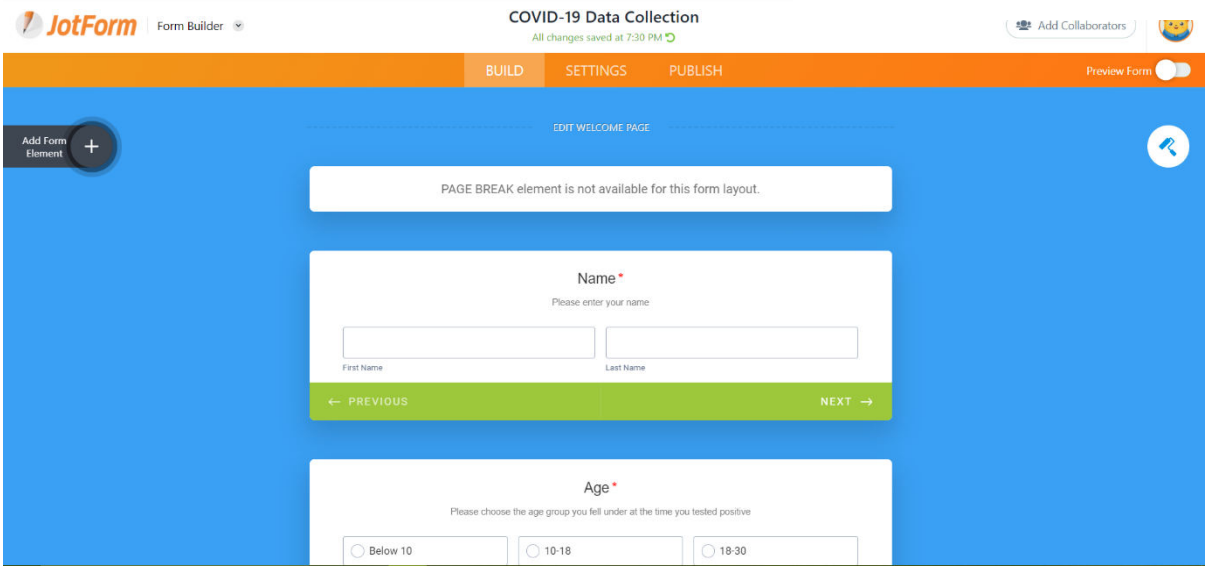


Fig.2 Form in editing mode

3. Distributing the form & collecting data

After the form was ready, the form was sent out to people who had tested positive. The people also gave feedback. Most of them appreciated the form, while some had complaints about the form's length. After everybody had filled the form, it had 57 submissions. These were saved in a JotForm table, the link for which is: https://www.jotform.com/tables/211304594729054.

Fig. 3 The JotForm table

4. Extracting data & creating charts

The data was then extracted from the JotForm table into a Google Sheet, for making charts. Care was taken so that the data doesn't get manipulated. After that, data was organized into charts *(refer to Discussion)* to cluster it. Trends were identified and noted*(refer to Results)*.

ISSN 2350-109X
www.indianscholar.co.in

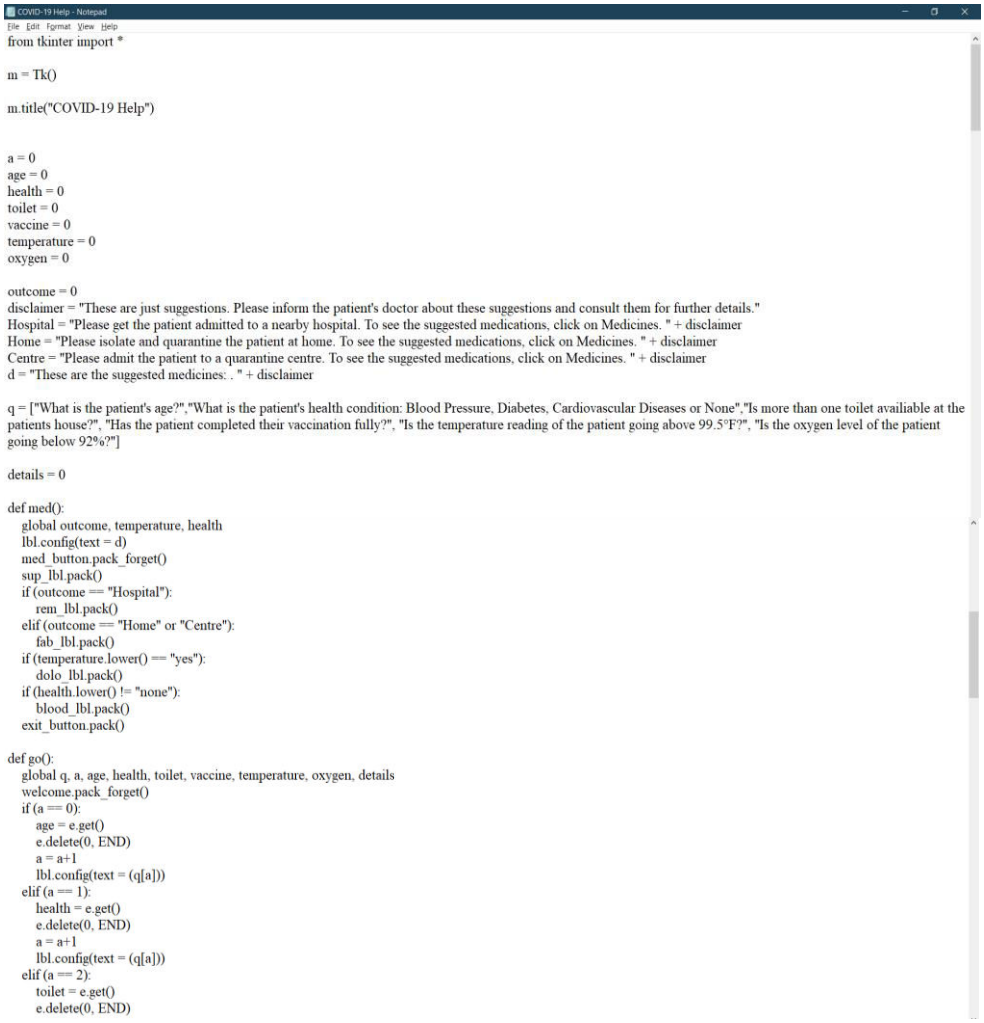
Indian Scholar

An International Multidisciplinary Research e-Journal

MISA
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

5. Creating an algorithm

Using the trends identified and some general background information, an algorithm was made in Python language (using Tkinter). This algorithm would ask for the patient's age, health condition, more than one toilet availability, vaccination, oxygen level, and temperature, and give them advice on the next steps. More conditions can be added to this algorithm, to accommodate more data points. It can also be expanded to an application, connected to the source table, which can be used by anyone.



```
COVID-19 Help - Notepad
File  Edit  Format  View  Help
from tkinter import *

m = Tk()

m.title("COVID-19 Help")


a = 0
age = 0
health = 0
toilet = 0
vaccine = 0
temperature = 0
oxygen = 0

outcome = 0
disclaimer = "These are just suggestions. Please inform the patient's doctor about these suggestions and consult them for further details."
Hospital = "Please get the patient admitted to a nearby hospital. To see the suggested medications, click on Medicines. " + disclaimer
Home = "Please isolate and quarantine the patient at home. To see the suggested medications, click on Medicines. " + disclaimer
Centre = "Please admit the patient to a quarantine centre. To see the suggested medications, click on Medicines. " + disclaimer
d = "These are the suggested medicines: . " + disclaimer

q = ["What is the patient's age?","What is the patient's health condition: Blood Pressure, Diabetes, Cardiovascular Diseases or None","Is more than one toilet availiable at the patients house?", "Has the patient completed their vaccination fully?", "Is the temperature reading of the patient going above 99.5°F?", "Is the oxygen level of the patient going below 92%?"]

details = 0

def med():
    global outcome, temperature, health
    lbl.config(text = d)
    med_button.pack_forget()
    sup_lbl.pack()
    if (outcome == "Hospital"):
        rem_lbl.pack()
    elif (outcome == "Home" or "Centre"):
        fab_lbl.pack()
    if (temperature.lower() == "yes"):
        dolo_lbl.pack()
    if (health.lower() != "none"):
        blood_lbl.pack()
    exit_button.pack()

def go():
    global q, a, age, health, toilet, vaccine, temperature, oxygen, details
    welcome.pack_forget()
    if (a == 0):
        age = e.get()
        e.delete(0, END)
        a = a+1
        lbl.config(text = (q[a]))
    elif (a == 1):
        health = e.get()
        e.delete(0, END)
        a = a+1
        lbl.config(text = (q[a]))
    elif (a == 2):
        toilet = e.get()
        e.delete(0, END)
```

Fig. 4a The algorithm

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

MISA
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

```
        a = a+1
        lbl.config(text = (q[a]))
    elif (a == 3):
        vaccine = e.get()
        e.delete(0, END)
        a = a+1
        lbl.config(text = (q[a]))
    elif (a == 4):
        temperature = e.get()
        e.delete(0, END)
        a = a+1
        lbl.config(text = (q[a]))
    elif (a == 5):
        oxygen = e.get()
        e.delete(0, END)
        a = a+1

    if (a == 6):
        e.pack_forget()
        welcome.config(text = "Please check these details. If they are incorrect please retry.")
        welcome.pack()
        details = "Age - " + str(age) + "   Health Condition - " + str(health) + "  Availability of more than one Toilet - " + str(toilet) + "  Vaccination completed - " + str(vaccine)
+ "   Temperature above 99.5° F - " + str(temperature) + "    Oxygen below 92% - " + str(oxygen)
        lbl.config(text = details)
        a = a+1
    if (a == 7):
        welcome.pack_forget()
        next_button.pack_forget()
        finish_button.pack()


def finish():
    global Hospital, Home, Centre, outcome, age, health, toilet, vaccine, temperature, oxygen, medicines
    e.pack_forget()
    next_button.pack_forget()
    finish_button.pack_forget()
    med_button.pack()

    if (int(age) < 60):
        if (health.lower() == "none"):
            if (oxygen.lower() == "no"):
                if (toilet.lower() == "yes"):
                    outcome = "Home"
                elif (toilet.lower() == "no"):
                    outcome = "Centre"
                else:
                    outcome = "Error"
            elif (oxygen.lower() == "yes"):
                outcome = "Hospital"
            else:
                outcome = "Error"
        elif (health.lower() != "none"):
            outcome = "Hospital"
        else:
            outcome = "Error"
    elif (int(age) >= 60):
        if (vaccine.lower() == "yes"):
            if (toilet.lower() == "yes"):
                outcome = "Home"
            elif (toilet.lower() == "no"):
                outcome = "Centre"
            else:
                outcome = "Error"
        elif (vaccine.lower() == "no"):
            outcome = "Hospital"
        else:
            outcome = "Error"
    else:
        outcome = "Error"

    if (outcome == "Hospital"):
        lbl.config(text = Hospital)
    elif (outcome == "Centre"):
        lbl.config(text = Centre)
    elif (outcome == "Home"):
        lbl.config(text = Home)
    else:
        lbl.config(text = "Error. Please retry.")


welcome = Label(m, text = "Welcome to COVID-19 help. This module will suggest a possible line of treatment for a COVID-19 positive patient. Please click finish only
after answering the sixth question.")
welcome.pack()

lbl = Label(m, text = (q[a]))
lbl.pack()
```

Fig. 4b The algorithm

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

MISA
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

```
sup_lbl = Label (m, text = "Vitamins and Minerals Supplements")
rem_lbl = Label (m, text = "Remdesivir")
fab_lbl = Label (m, text = "Fabiflu")
dolo_lbl = Label (m, text = "Dolo")
blood_lbl = Label (m, text = "Blood Thinners")

e = Entry(m)
e.pack()


next_button = Button(m, text = "Next", fg = "black", bg = "white", height = 1, width = 7, command = lambda:go())
next_button.pack()

finish_button = Button(m,text = " Finish", fg = "black", bg = "white", height = 1, width = 7, command = lambda:finish())

med_button = Button(m, text = "Medicines", fg = "black", bg = "white", height = 1, width = 7, command = lambda:med())

exit_button = Button(m, text = "Exit", fg = "black", bg = "white", height = 1, width = 7, command = m.destroy)


m.mainloop()
```

Windows (CRLF) Ln 115, Col 33 100%

Fig. 4c The algorithm

6. Testing & Classification

   The algorithm was tested and run. Dummy patient details were entered as a test. The algorithm gave the correct suggestion.

RESULT

The experiment was carried out to check the credibility of the suggested methodology. The experiment was successful and proved the methodology credible. This methodology can very well be carried out by data scientists in an epidemic/pandemic. A little help from doctors/medical staff will increase the quality of the results.

DISCUSSION

The following charts (Fig. 5a – 5f) show how scores of COVID-19 related tests change with age. Clusters of RT-PCR scores are falling as age increases which denote the severity of infections are rising. Clusters of CRP and HRCT scores are rising with age, which also denotes the same.

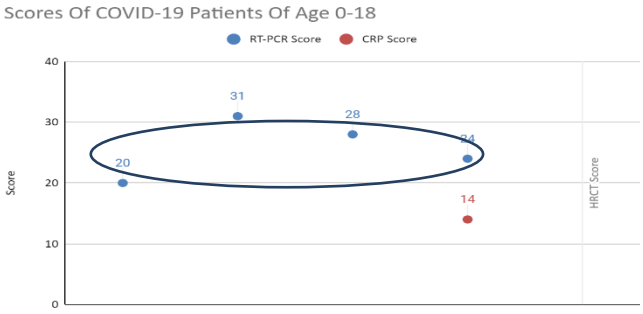


Fig. 5a Age vs Scores

Fig. 6 shows how health condition affects your final

ISSN 2350-109X
www.indianscholar.co.in
Indian Scholar
An International Multidisciplinary Research e-Journal
MISA Global Partnership
MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

outcome. If the patient has a health condition, it increases their chances of being hospitalized. Fig. 7 shows the relationship between age and days infected; days infected increase with age. "As age increases, the chances of you being hospitalized increases", this trend was clearly shown by Fig. 8. Fig. 9 shows what medication was given to each age group.

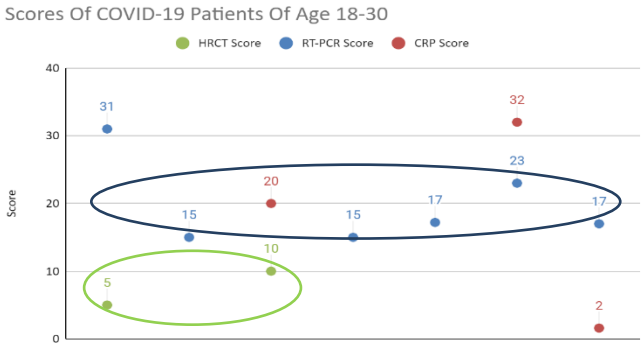


Fig. 5b Age vs Scores

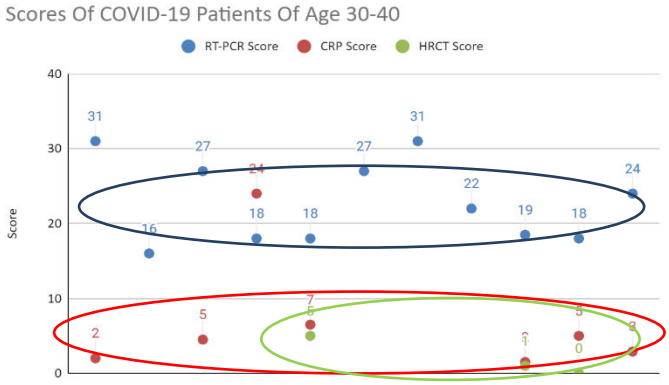


Fig. 5c Age vs Scores

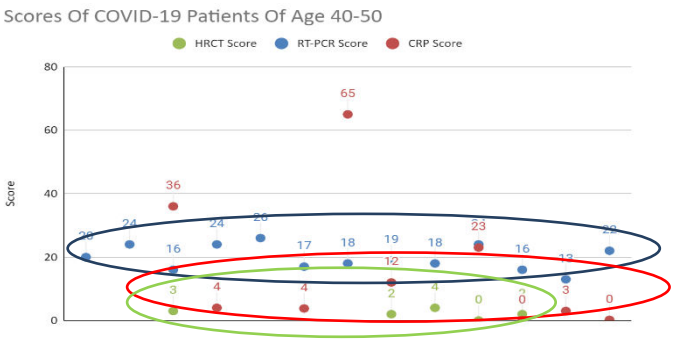


Fig. 5d Age vs Scores

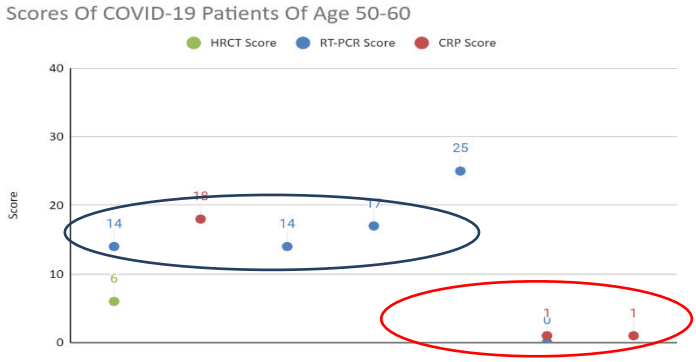


Fig. 5e Age vs Scores

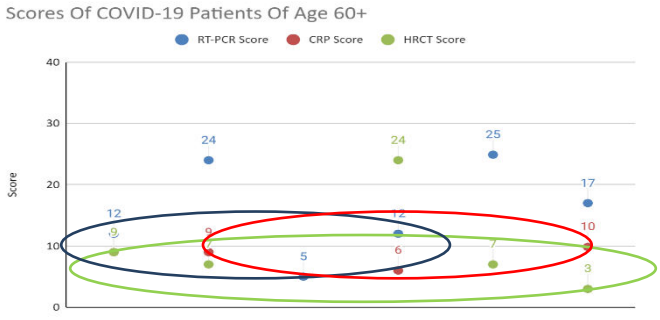


Fig. 5f Age vs Scores

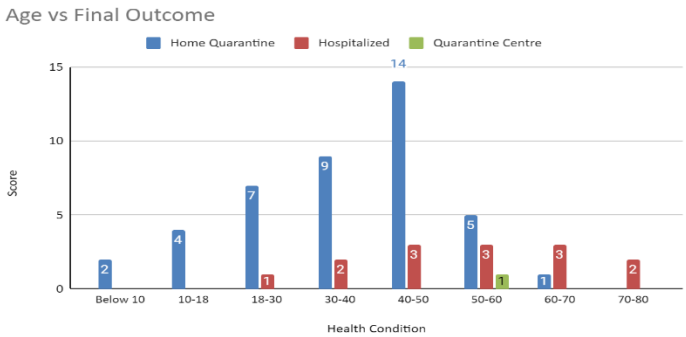


Fig. 8 Age vs Final Outcome

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

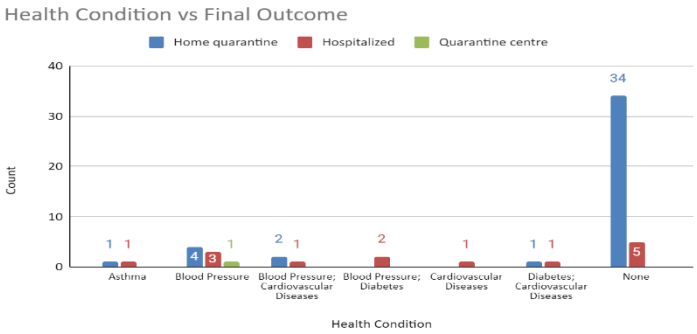
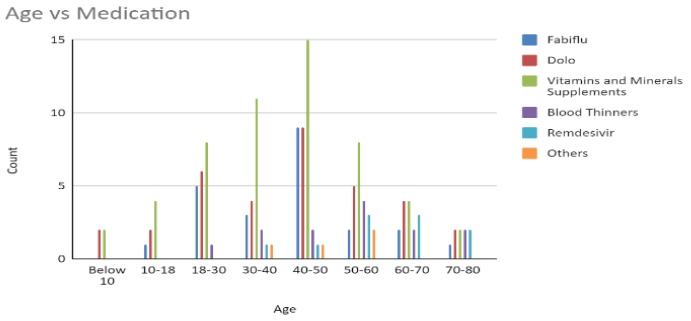
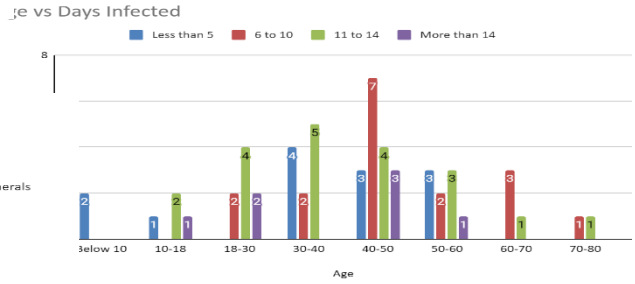


Fig. 6 Health Condition vs Final Outcome





ge vs Days Infected

Fig. 9 Age vs Medication

CONCLUSION

In conclusion, an experiment conducted on a small dataset of patients who recently contracted COVID-19 resulted in interesting findings. We were able to identify high-risk groups and their possible treatment methods all by clustering and classifying their symptoms, medication and test result values. Although, it has to be noted that every technique applied in Data Science requires constant and repetitive trials to attain a data set that could factually result in conclusions. 'Trial and Error'are probably the primary keywords of achieving any sort of success in the methods applied to gathering, filtering and making sense of the vast data in any industry.Gathering a vast number of structured data points on initial symptoms, test results, treatment plans and results of any epidemic and clustering and thereby classifying patients into various buckets of categories could turn out to be a game-changer in solving a future health crisis of epic proportions. We could use machines to do what they do best – learn through piles of health data and help conclude to solve humanity's medical problems.

Acknowledgements

Vikram Ramchand, Amudha Justin Manickam, Shiksha Tahiliani, Girish Tahiliani, everyone who filled the form.

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

REFERENCES

1. "Data Science." Data Robot.https://www.datarobot.com/wiki/data-science/(accessed Jun. 1, 2021).
2. M. Cárdenas "How Data Science can help in a pandemic situation?" Sopra Steria. https://www.soprasteria.com/insights/details/how-can-data-science-help-in-a-pandemic-situation(accessed Jun. 1, 2021).
3. "Clustering in MachineLearning." Geeks For Geeks. https://www.geeksforgeeks.org/clustering-in-machine-learning/(accessed Jun. 1, 2021).
4. "Data Mining - Cluster Analysis" Tutorialspoint https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm(accessed Jun. 1, 2021).
5. C. Davidson. "What Is Classification Analysis?" Indicative. https://www.indicative.com/data-defined/classification-analysis/(accessed Jun. 1, 2021).
6. J. Smith. "Quotes of the Week: John Tukey." DATA SCIENTIST INSIGHTS.https://datascientistinsights.com/2013/01/29/quotes-of-the-week-john-tukey/ (accessed Jun. 1, 2021)