IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

COVID-19 Survey: Data Analytics and Predictive Analysis

Dr Rama Chikkamuniswamy[1], Dr H S Manjula[2], Vishesh S[3], Suraj S[4], Rishi Singh[5]

MD, DM, Assistant Professor, Cardiology, SJICS&R, Bangalore, India [1]

MD, Assistant Professor, Biochemistry, SJICS&R, Bangalore, India [2]

BE (TCE), MBA (e-business) [3]

BE, ECE, SJBIT, Bangalore, India [4]

Student, BE, Mechanical, PES (South Campus), Bangalore, India [5]

Abstract: Coronaviruses are a group of viruses which have been said to have originated from Wuhan, China belonging to the family of Coronaviridae. Human coronaviruses can cause lung infections which can be fatal if left untreated. COVID-19 death is defined for surveillance purposes as a death resulting from a clinically compatible illness in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be related to COVID disease.[1] There should be no period of complete recovery between the illness and death. In our paper we have published the results based on a survey conducted by our team and, used Data Analytic tools and Predictive Analysis on the acquired data. Vital questions have been asked and opinions have been collected from around 500 residents of each area in Bangalore Urban zone. Visualization tools using Python libraries have refined our data visualization process. Sampling rate is fixed not to overfit or under fit during supervised learning using Artificial Intelligence (AI).

Keywords: Coronaviruses, Coronaviridae, COVID-19, Data Analytic tools and Predictive Analysis, Visualization tools, Python libraries, supervised learning using Artificial Intelligence (AI), Real- time PCR and CT images/ X-rays.

I.INTRODUCTION

The general symptoms of COVID-19 patients are flu-like such as fever, cough, dyspnoea, breathing problem, and viral pneumonia. But these symptoms alone are not significant. There are many cases where individuals are asymptomatic but their chest CT scan and the pathogenic test were COVID-19 positive. So, along with symptoms, positive pathogenic testing and positive CT images/X-Rays of the chest are being used to diagnose the disease. For pathological testing, Real- time PCR is being used as a standard diagnostic tool. Healthcare systems around the world are attempting to expand testing facilities for COVID-19.[2] More and more testing will lead to the identification and isolation of infected persons, thereby reducing the spread among the community. But availability does not ensure reliability. The major concern for the governments at this stage is the false negative test results – the test results are negative for the infected individual. Such individuals may unknowingly transmit the virus to others. False test results thus have a negative effect on the efforts to curb the spread of the virus.[3] The impact of this concern on the safety of public and health workers can't be determined as there are no clear or consistent reports on these test performance characteristics. The sensitivity of these tests is largely unknown.[4][5][6]

II.DATA PRE-PROCESSING

COVID-19 is a global pandemic and the survey has resulted in huge amounts of data. Data is being stored, exchanged and conditioned. The volume of data that one has to deal with has exploded to unimaginable levels. Most of the data exists in its crude form and needs to be converted to useful format before analysis. This process of converting raw data into useful format is called data pre-processing. Real world data is

- Incomplete: consists of missing attribute values or consists of only aggregate data
- Noisy: containing errors or outliers.
- Inconsistent: containing discrepancies in code.
- May contain non numerical values
- May be Redundant
- May not fit to scale

Figure 1 shows the process of data pre-processing and data preparation for visualization using Python and its repositories. Raw data needs to be structured before exploration and visualization. There are many reasons why data might be missing in a dataset. For instance, data collected through a survey may have missing data due to participants'

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

failure to respond to some questions, not knowing the correct response, or being unwilling to answer. It may also be missing due to the error made during the data entry process.

Figure 1 shows the process of data pre-processing and data preparation for visualization.

III. SURVEY QUESTIONS

1. Does Govt need any change in services to ensure public safety and health at this time?
2. Should current update of bed availability in all the hospitals be published on social media?
3. Proper COVID-19 testing is done in your current area?
4. Can Govt start public transportation like trains and metro?
5. Should Govt provide more number of Ambulances only for COVID-19 patients?
6. Are their sufficient Quarantine centres or hospitals?
7. Can Govt start schools, colleges and IT sectors during COVID-19?
8. Proper medication and testing kits are supplied to hospitals?
9. Should Gym, Pubs and Movie Theatres remain closed for another 2 months for maintaining social distancing?
10. Did you ever have any symptoms of COVID-19?
11. You are/ were you a COVID-19 patient?
12. Have you been quarantined in any Govt quarantine centre?
13. Do you agree on imposition of another lockdown?
14. Do you agree that the present Govt services provided during COVID-19 is enough or would you demand for improved services and aid?
15. Is or was your area in a containment zone ever?
16. Please enter your area PINCODE.

IV. DATA VISUALIZATION

In the previous section raw data was cleansed and structured. The pre-processed time stamped data is as shown in figure 2. The pre-processed data needs to be checked for Missing values or redundancy before Exploration Data Analysis (EDA) and data visualization. Figure 3 shows the code to check for missing values or redundant values. The columns must not contain NaN (Not-a-Number) values. Each attribute of the medical survey data is checked for NaN values and refined by using 'fillna' statements. The opinion of each resident of an area with a unique PINCODE is noted to be either 'yes' (logic 1) or 'no' (logic 0). Boolean mapping is done with python programming language and a unique key is created based on the area. The survey conducted on COVID-19 is a huge task and involves exploding levels of data, and needs scale down procedure during visualization.[7][8] This can be achieved by data normalization techniques and formulas. Post Normalization the data would either lie between 0 and 1. Figure 4 shows the output of normalization. Correlation is an important data visualization entity. One can easily know the dependencies of each attribute with the other. It is important to discover and quantify the degree to which variables in your dataset are dependent upon each other. This knowledge can help you better prepare your data to meet the expectations of machine learning algorithms, such as linear regression and other supervised learning algorithms, whose performance will degrade with the presence of these interdependencies. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated. Figure 5 shows the heatmap of the process of correlation.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

In [3]: df

Out[3]:

	Timestamp	Do Govt need any change in services to ensure public safety and health at this time ?	Current update of beds availability in all hospitals should be given in social media ?	Proper survey of Covid-19 test check ups is done in your current area ?	Can Govt start the public transportation like train and metro ?	Should Govt has to provide more number of Ambulances only for Covid-19 patients ?	Should Govt has to keep Quarantine centers or Hospitals are sufficient ?	Can Govt start schools and colleges and other IT sectors in Covid-19 ?	Proper medication and testing kits are supplied to hospital ?	Should Gym Pubs and Movie Theaters should be remained closed for another 2 month for maintaining social distancing	Did you ever had any symptoms of Covid-19 ?	You are or Were a Covid-19 patient ?	have you been quarantined in any Govt quarantine centers ?	d... Lo...
0	6-30-2020 22:50:46	1	1.0	1.0	1	0.0	1	1.0	1.0	1	0.0	0.0	0.0	
1	6-30-2020 22:01:48	1	1.0	1.0	0	1.0	2	0.0	0.0	0	0.0	0.0	0.0	
2	6-30-2020 22:02:03	1	1.0	0.0	1	1.0	2	1.0	0.0	1	0.0	0.0	0.0	
3	6-30-2020 22:02:57	1	1.0	1.0	0	1.0	1	0.0	0.0	1	0.0	0.0	0.0	
4	6-30-2020 22:04:28	1	1.0	1.0	0	1.0	2	0.0	0.0	1	1.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
318	09-04-2020 20:53	1	0.0	0.0	0	1.0	0	0.0	0.0	0	0.0	0.0	0.0	
319	09-04-2020 20:53	1	0.0	1.0	0	1.0	0	0.0	0.0	0	1.0	1.0	0.0	
320	09-04-2020 20:54	0	1.0	1.0	1	0.0	2	1.0	0.0	0	1.0	0.0	1.0	
321	09-04-2020 20:54	1	1.0	0.0	0	1.0	2	0.0	1.0	1	1.0	1.0	1.0	
322	09-04-2020 20:54	1	1.0	1.0	0	1.0	2	0.0	0.0	1	0.0	0.0	0.0	

Figure 2 shows time stamped pre-processed data.

```
In [8]: df.isnull().any()

Out[8]: Timestamp                                                                                  False
        Do Govt need any change in services to ensure public safety and health at this time ?      False
        Current update of beds availability in all hospitals  should be given in social media ?     True
        Proper survey of Covid-19 test check ups is done in your current area ?                     True
        Can Govt start the public transportation like train and metro ?                            False
        Should Govt has to provide more number of Ambulances only for Covid-19 patients ?           True
        Should Govt has to keep Quarantine centers or Hospitals are sufficient ?                   False
        Can Govt start schools and colleges and other IT sectors in Covid-19 ?                      True
        Proper medication and testing kits are supplied to hospital ?                               True
        Should Gym Pubs and Movie Theaters should be remained closed for another 2 month for maintaining social distancing ?  False
        Did you ever had any symptoms of Covid-19 ?                                                 True
        You are or Were a Covid-19 patient ?                                                        True
        have you been quarantined in any Govt quarantine centers ?                                  True
        Do you agree on making of Lockdown 6.0  ?                                                  False
        Your area is or was in a containment zone ever ?                                           False
        age                                                                                         True
        bloodpressure                                                                               True
        dtype: bool

In [9]: df.isnull().sum()

Out[9]: Timestamp                                                                                   0
        Do Govt need any change in services to ensure public safety and health at this time ?       0
        Current update of beds availability in all hospitals  should be given in social media ?     1
        Proper survey of Covid-19 test check ups is done in your current area ?                     2
        Can Govt start the public transportation like train and metro ?                             0
        Should Govt has to provide more number of Ambulances only for Covid-19 patients ?           5
        Should Govt has to keep Quarantine centers or Hospitals are sufficient ?                    0
        Can Govt start schools and colleges and other IT sectors in Covid-19 ?                      1
        Proper medication and testing kits are supplied to hospital ?                               9
        Should Gym Pubs and Movie Theaters should be remained closed for another 2 month for maintaining social distancing ?  0
        Did you ever had any symptoms of Covid-19 ?                                                 1
        You are or Were a Covid-19 patient ?                                                        1
        have you been quarantined in any Govt quarantine centers ?                                  1
        Do you agree on making of Lockdown 6.0  ?                                                   0
        Your area is or was in a containment zone ever ?                                            0
```

Figure 3 shows the code to check for missing values or redundant values.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

```
In [36]: fsplt.subplots(figsize=(10,10))
corr =df.corr()
corr= (corr)
sns.heatmap(corr,xticklabels=corr.columns.values,linewidth=0.5,
yticklabels=corr.columns.values)
corr
```

Out[36]:

	sino	change	bed	test	transportation	Ambulances	sufficient	schools	medication	Theaters	symptoms	Covid-19	quar
sino	1.000000	-0.280890	-0.236780	-0.022005	0.061660	-0.336899	-0.069423	0.258175	0.209100	0.024204	0.377662	0.313030	0
change	-0.280890	1.000000	0.277619	-0.054049	-0.092396	0.324968	0.047245	-0.124528	-0.086675	0.151550	-0.098637	-0.063758	-0
bed	-0.236780	0.277619	1.000000	-0.076948	-0.136904	0.216036	0.038465	-0.171961	-0.024750	0.140329	-0.131176	-0.092530	-0
test	-0.022005	-0.054049	-0.076948	1.000000	0.197149	-0.069676	-0.022849	0.279788	0.214911	-0.015605	0.098260	0.201146	0
transportation	0.061660	-0.092396	-0.136904	0.197149	1.000000	-0.125773	-0.047788	0.408302	0.137018	-0.025436	0.180945	0.174000	0
Ambulances	-0.336899	0.324968	0.216036	-0.069676	-0.125773	1.000000	0.061028	-0.138966	-0.005161	0.145272	-0.091680	-0.137358	-0
sufficient	-0.069423	0.047245	0.038465	-0.022849	-0.047788	0.061028	1.000000	0.010066	-0.019321	0.161269	-0.004581	0.074805	-0
schools	0.258175	-0.124528	-0.171961	0.279788	0.408302	-0.138966	0.010066	1.000000	0.314587	-0.092159	0.427008	0.354385	0
medication	0.209100	-0.086675	-0.024750	0.214911	0.137018	-0.005161	-0.019321	0.314587	1.000000	0.027402	0.211782	0.236189	0
Theaters	0.024204	0.151550	0.140329	-0.015605	-0.025436	0.145272	0.161269	-0.092159	0.027402	1.000000	-0.028738	0.058978	0
symptoms	0.377662	-0.098637	-0.131176	0.098260	0.180945	-0.091680	-0.004581	0.427008	0.211782	-0.028738	1.000000	0.423950	0
Covid-19	0.313030	-0.063758	-0.092530	0.201146	0.174000	-0.137358	0.074805	0.354385	0.236189	0.058978	0.423950	1.000000	0

Figure 4 shows the output of normalization.

V. OBSERVATIONS

i. More than 75% of the subjects need better services to ensure public safety.
ii. More than 88% of the subjects are complaining of bed unavailability or shortage of beds in Govt and private hospitals.
iii. Approximately 40% of the subjects are for reopening of Schools, Colleges and IT sectors?
iv. More than 90% of the subjects agreed that they were tested for COVID-19.
v. Around 20% of the subjects are staying in containment zones.
vi. Less than 10% agree that Gyms and Theatres must be opened for the public.
vii. Around 35% people are demanding for public transports for their convenience and daily needs.

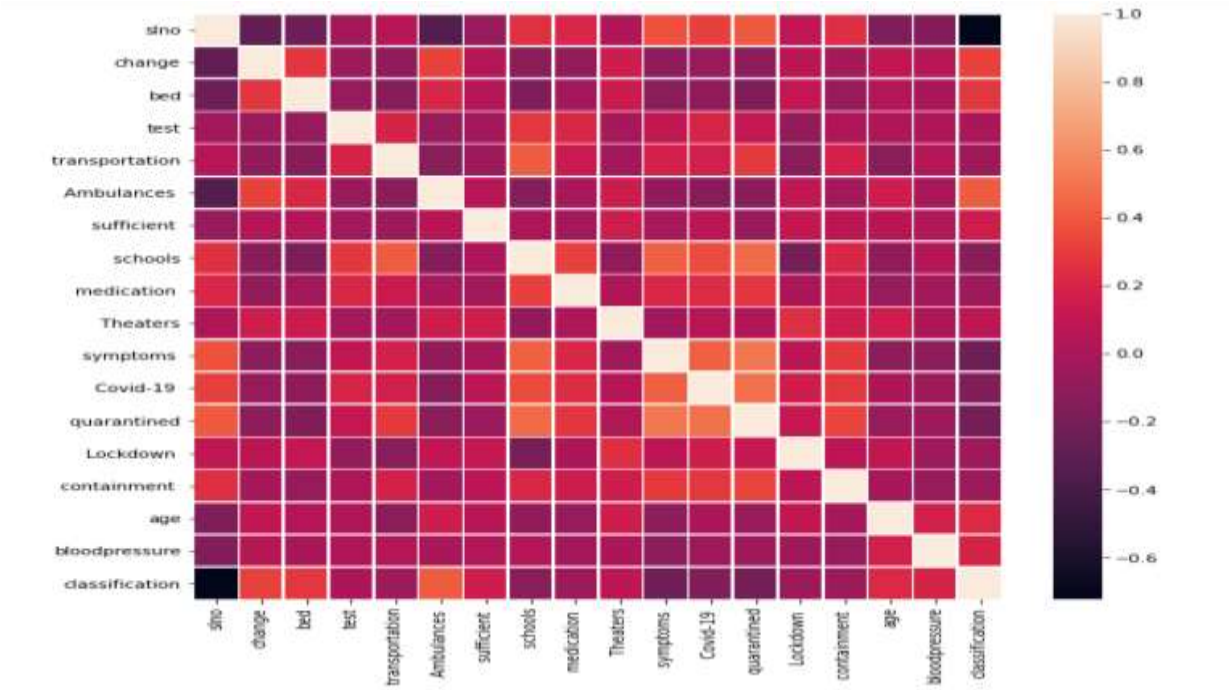

Figure 5 shows the heatmap of the process of correlation.

VI. PREDICTIVE ANALYSIS

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

We have collected opinions of 500 residents in each area. Using sampling techniques we target the entire community or population of the area. Using the collected data prediction can be done for the entire area using supervised machine learning techniques (Predictive Analysis). We have chosen two ML algorithms, Logistic Regression and Random Forest Classifier technique of binary classification. Figure 6 and 7 shows the execution of the ML model on the collected dataset. Figure 8 shows the verification of the designed model. We have inspected the accuracy, precision, recall, and F1 score of both Logistic Regression and Random Forest Classifier models. ROC plot is used to compare the accuracies of both the models with the base prediction model, to select the best fit for prediction. Figure 9 shows the ROC plot of the assignment.

```
from sklearn.linear_model import LogisticRegression


print ("---Base Model---")
base_roc_auc = roc_auc_score(y_test, base_rate_model(X_test))
print ("Base Rate AUC = %2.2f" % base_roc_auc)
print(classification_report(y_test, base_rate_model(X_test)))

# NOTE: By adding in "class_weight = balanced", the Logistic Auc increased by about 10%! This adju
sts the threshold value
logis = LogisticRegression(class_weight = "balanced")
logis.fit(X_train, y_train)
print ("\n\n ---Logistic Model---")
logit_roc_auc = roc_auc_score(y_test, logis.predict(X_test))
print ("Logistic AUC = %2.2f" % logit_roc_auc)
print(classification_report(y_test, logis.predict(X_test)))
```

Figure 6 shows the execution of the ML model on the collected dataset (Logistic Regression).

```
# Random Forest Model
rf = RandomForestClassifier(
    n_estimators=1000,
    max_depth=None,
    min_samples_split=10,
    class_weight="balanced"
    #min_weight_fraction_leaf=0.02
    )
rf.fit(X_train, y_train)
print ("\n\n ---Random Forest Model---")
rf_roc_auc = roc_auc_score(y_test, rf.predict(X_test))
print ("Random Forest AUC = %2.2f" % rf_roc_auc)
print(classification_report(y_test, rf.predict(X_test)))
```

Figure 7 shows the execution of the ML model on the collected dataset (Random Forest Classifier).

```
 ---Logistic Model---
Logistic AUC = 0.74
             precision    recall  f1-score   support

          0       0.90      0.76      0.82      1714
          1       0.48      0.73      0.58       536

avg / total       0.80      0.75      0.76      2250


 ---Random Forest Model---
Random Forest AUC = 0.97
             precision    recall  f1-score   support

          0       0.99      0.98      0.99      1714
          1       0.95      0.96      0.95       536

avg / total       0.98      0.98      0.98      2250
```

Figure 8 shows the verification of the designed model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101008

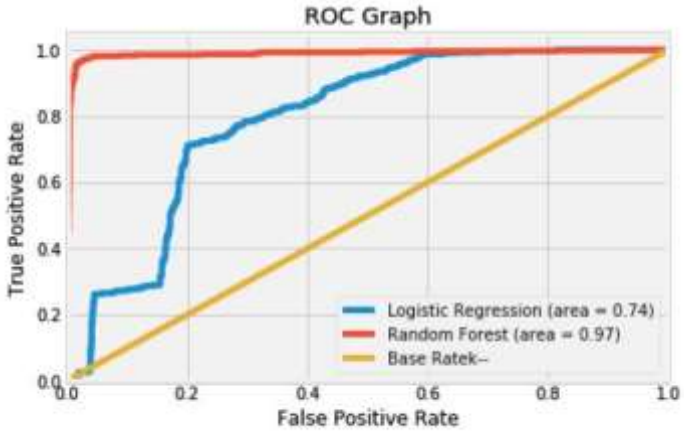

Figure 9 shows the ROC plot of the assignment.

VII. CONCLUSIONS

An opinion survey was conducted in real time, samples obtained and stored in a server for real time analysis, and future predictions regarding COVID-19. This survey is an exact measure of how the virus is behaving and its spread in a community. Also conclusions can be drawn on how an individual and community react to such a wide and quick spreading pandemic like COVID-19. The dataset can be used as a benchmark for future attacks by mutated cornavirues. Data analytics had to be carried out on the data –both historical and present trend to draw inference. The goal was to create a database of smaller sample size- 500 samples from each area belonging to Bangalore Urban district and predict the opinions of the entire community or area by predictive analysis. A python code was written and executed to analyse and draw conclusions. The first step in data analytics- data pre-processing was successfully carried out and a heatmap was plotted to inspect any interdependencies. ROC graph which can easily indicate the difference in the performance of the 2 predictive models was plotted. The accuracy, precession, and F1 score was obtained and compared. The Accuracy of Random Forest Classifier and Logistic Regression were found to be nearly 97% and 74% respectively. Random Forest is the best fit for this assignment.

REFERENCES

[1] INTERNATIONAL GUIDELINES FOR CERTIFICATION AND CLASSIFICATION (CODING) OF COVID-19 AS CAUSE OF DEATH. Based on ICD International Statistical Classification of Diseases (16 April 2020).

[2] P. Chatterjee et al., "The 2019 novel coronavirus disease (Covid-19) pandemic: A review of the current evidence," Indian Journal of Medical Research, Supplement, vol. 151, no. 2–3. Indian Council of Medical Research, pp. 147–159, 2020, doi: 10.4103/ijmr.IJMR_519_20.

[3] G. Pascarella et al., "COVID-19 diagnosis and management: a comprehensive review," Journal of Internal Medicine, 2020, doi: 10.1111/joim.13091.

[4] C. P. West, V. M. Montori, and P. Sampathkumar, "COVID-19 Testing: The Threat of False- Negative Results," Mayo Clinic Proceedings. Elsevier Ltd, 2020, doi: 10.1016/j.mayocp.2020.04.004.

[5] D. Wang et al., "Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China," JAMA - Journal of the American Medical Association, vol. 323, no. 11, pp. 1061–1069, Mar. 2020, doi: 10.1001/jama.2020.1585.

[6] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," The Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140- 6736(20)30183-5.

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1207–1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[8] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.05225.

