Towards Understanding the Psychological Effects of the COVID-19 Pandemic on the Indian Population

Deepanshu Pandey*
ZS Associates India PVT. LTD.
New Delhi, India
deepanshupandey195@gmail.com

Pranav Khurana*
Netaji Subhas University of Technology
New Delhi, India
pranavkhurana24@gmail.com

Ashwin Misra*
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, United States
amisra@andrew.cmu.edu

***Abstract*—This article explains the preliminary results of the analysis of a public survey carried out in India, assessing the psychological effects on people during the second wave of the COVID-19 pandemic. A survey was designed to categorize the population on the basis of various socio-economic demographics and respondents were then asked to fill out the DASS-21 questionnaire to get their levels of severity of anxiety, depression and stress. The dataset obtained was then further analyzed using various classification machine learning models with the level of severity as the target variable and respondent's attributes as independent variables. A Multinomial Logistic Regression was found to give the best results with an AUC score of 0.94 and was thus, used to predict the severity levels of these three categories, to find various insights from this publicly-sourced dataset. Additionally, the significance of the various socio-demographic attributes asked in the survey was analyzed in order to identify key drivers of mental ailments among the general Indian population. Further, a brief description of segmenting the population using K-Means clustering is provided which attempts to identify population groups that belong to similar socio-economic demographics and suffer from similar mental health issues during the pandemic. Thus, high-risk or high-severity groups can be identified and then could be targeted by the government to provide them relief schemes. This paper applies machine learning on a public dataset to explore the various facets of COVID-induced problems in the Indian Society.**

***Index Terms*—COVID-19 Pandemic, Multinomial Logistic Regression, Clustering, Population Segmentation**

I. Introduction

The second wave of COVID-19 in India started in early April 2021 when there was a huge spike in number of the daily new cases and a highly unprecedented infectious coronavirus variant, B.1.617 was also identified. In early-May 2021, the number for new cases per day were above 350,000 (the greatest number of cases in a single day in 2020 were 98,000) and the number of deaths per day were above 3,500 which led to another nationwide lockdown. According to a study [1], one in seven Indians have mental disorders of varying severity and when combined with a huge shortage of medical facilities (ICU beds, testing and isolation centers, oxygen, etc.), crematoriums, and burial grounds during the second wave caused even more mental health challenges to the citizens of India and increased the intensity of negative psychological outcomes. Anxiety, depression, and post-traumatic stress were among the most common problems reported by callers to the government's COVID mental health helpline during these times. [2].

A study [3] conducted during the 2020's lockdowns in India found that the lockdowns caused major disruptions to daily routines of the people and hindered their ability to meet regular responsibilities. This ultimately affected the physical and mental health of individuals and people experienced increased psychopathological symptoms during COVID-19 outbreak as compared to pre-COVID-19 times [4]. India is said to be the world's most depressing country and it spends less than 2% of its annual health budget on mental health [5]. Many Indians still view mental health as a social stigma and globally 20% of young people experience mental disorders but in India, only 7.3% of its 365 million youth report such problems and about one-third of young people display poor knowledge of mental health problems and negative attitudes towards people with mental health problems and one in five had actual/intended stigmatizing behavior [6].

A survey by World Health Organization (WHO), found that COVID-19 pandemic has disrupted critical mental health services in 93% of countries worldwide while the demand for mental health is increasing [7]. Therefore, in a country like India (which is the world's second most populated country) with a rich social and cultural diversity, there is a dire need to solve country's mental health issues especially during devastating times of COVID-19.

With this article, we aim to find the key socio-economic and behavioral attributes of the Indian population that cause severe mental health issues during the pandemic. The data set is created through a thorough survey which imparts high confidence and reliability to the data. Supervised machine learning, used for classification or prediction modelling, has the advantage of accounting for complex relationships between variables that may not have been previously identified and show great promise in improving the diagnosis and treatment of patients with mental health conditions [8].

Therefore, a preliminary approach to predict the severity levels of anxiety, depression and stress of individuals in the country is suggested. The data gives us several insights into the various psychological aspects during the pandemic, which can not be obtained by directly asking targeted questions as the Indian populace is still reluctant and stigmatic when it comes to mental health. If the high-risk groups can be

*Authors contrinuted equally

identified, it will ultimately help in decreasing the depression rate by extending help and providing resources to those who are suffering with severe mental health issues. Therefore, a clustering approach to segment the population is also proposed which would help in identifying the parts of the society vulnerable to serious mental health issues. The main idea behind first predicting, and then clustering, is that prevention is better than cure.

II. DATA COLLECTION AND MODELLING METHODOLOGY

A. Background and Questionnaire Description

Previous works in the domain of analyzing the general mental health of Indian citizens has been limited in scope, [9] used a Random Forest classifier to predict anxiety, depression and stress using data collected from a survey using the DASS-21 Questionnaire. However, the results were not interpretable and the model didn't employ any behavioral or social characteristics of the respondents as features. The article [10] conducted a similar study but it was limited to a very specific demographic and it did not dive deep into the key drivers of the severity of mental health issues. This article focuses on identifying varying levels of anxiety, depression, and stress in a person based on their socio-economic status, and the impact of COVID-19 in their lives, as well as the behavioral characteristics during the lockdown period.

A total of 495 responses were collected from a Google survey which was circulated in various social media platforms which asked questions in these three verticals:

• Socio-economic characteristics: Age, occupation, gender, annual household income (in Rupees), number of earning members in the family, pending loan(s)

• Behavioral and COVID-19 related characteristics: whether or not infected by COVID-19, whether received any COVID-19 vaccine, health comorbidities, impact of COVID-19 in their lives, income change due to COVID-19, number of times exercise/yoga done in a week, number of times socially interacted with family/friends in a week, change in sleep or appetite during lockdown

• Depression, Anxiety, Stress Scale (DASS-21) questionnaire

DASS-21 [11] essentially consists of 21 questions, with 7 questions allocated to anxiety, depression and stress each and it was used to gauge the severity level of anxiety, depression, and stress level of a particular respondent. The four severity levels were: (i) Normal (ii) Mild (iii) Moderate and (iv) Severe.

For a population size of 1.3 billion (India's population), confidence interval of 95% and margin of error of 5%, the sample size requirement is 385 [12] which is well below the total responses received from the Google survey i.e., n=495.

B. Data Preprocessing

The data received from the survey was processed in order to make it suitable for classification machine learning models. The following techniques were applied in Python 3.8:

1) Age was categorized into 4 categories: 16-18 years old, 18-25 years old, 25-45 years old, 45+ years old and similarly, Annual Household Income (in Rupees) was categorized in 4 categories: less than 5 lakhs, 5-10 lakhs, 10-20 lakhs, 20+ lakhs
2) Gender, Occupation, Health comorbidities were one hot encoded and rest of the attributes were label encoded ordinally
3) A train-test split of 70:30 was selected and the dependent variables were level of severity of anxiety/depression/stress as calculated from DASS-21 (i.e., Normal, Mild, Moderate or Severe)

C. Data Exploration

The 495 responses received for the survey [13] had an even mix of respondents across all levels of socio-demographic characteristics as asked in the questionnaire. The subjects were chosen from a wide spectrum of the population covering different ages, income groups aw well as other parameters.

Figure 1. depicts the split of respondents across various employment backgrounds. As it can be seen that the most employed portion is in the private sector followed by the Unemployed sector. This points to the rise in the unemployment level due to the pandemic which is probably the most important driving factor for this study. *Figure 3.* depicts the split of respondents across various levels of income change as experienced by their families over the past couple of months. It can be inferred that most of the subjects didn't see a fluctuation in their incomes. The second largest group is which noticed a slight decrease which should not generally amount to socio-psychological effects. The blue region of significant decrease is most probably the region where distress and depression can be most observed. There is also an increase in salary level seen which may be due to different company policies, type of job, and in-person requirements etc. *Figure 2.* depicts the split of respondents across various levels of change in appetite as observed by them over the past couple of months. If a direct correlation is drawn from the decrease in salary, there should be a decrease in appetite due to the low purchasing power of the subjects, although, the significant portion highlights a no-change portion. Though the second major region is a drop in appetite, which is logical with the purchasing power reduction, although the slight increase group is also formidable. This can be due to lethargy and the lack of physical activities and the adoption of a sedentary lifestyle. Both of the groups pointing to a possible change in the psychological state of the subjects.

D. Modelling

Three separate models (one each for Anxiety, Depression and Stress) with respondent's attributes as independent variables and the level of anxiety/depression/stress (as obtained from DASS-21 questionnaire) as the dependent

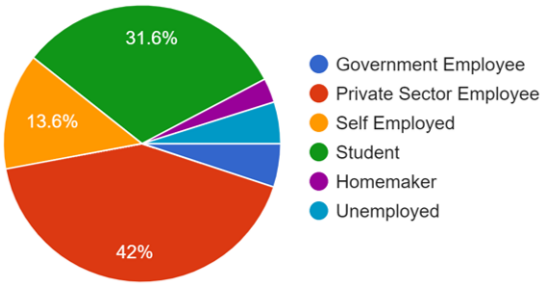

Fig. 1. Split across Employment backgrounds

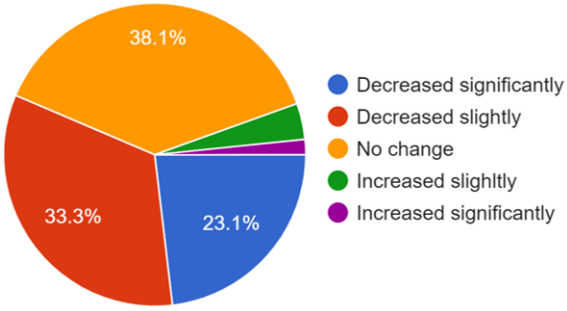

Fig. 2. Split across changes in Income

variable were tried and tested using different classification models. GridSearchCV, a function from python's sklearn library was used to find the ideal set of hyperparameters which optimizes the cost function and gives the best possible accuracy. Following is an example of best set of hyperparameters obtained using GridSearchCV for 4 different classification models for anxiety:

(i) **Logistic Regression** C: "1", solver: 'lbfgs', multi_class='multinomial'
(ii) **Support Vector Machines** C: 10, decision_function_shape: 'ovo', gamma: 'scale', kernel: 'rbf'
(iii) **Decision Tree** criterion: 'entropy', max_depth: 24, min_sample_leaf': 5, min_samples_split: 7, splitter: 'random'
(iv) **Random Forest** criterion: 'entropy', max_depth: 10, max_features: 9, n_estimators: 20

Following is a brief explanation of the hyper-parameters involved:

1) **C:** Represents the inverse of the regularisation strength. This parameters is adjusted in order to obtain a balance between bias and variance. With various tries, 1-10 was the most suited for our dataset.
2) **multi_class:** This parameter is used to specify the number of classes in the output. Multinomial means that the output class is not binary but instead has multiple classes. In the public dataset obtained, there is amultiple classes case.

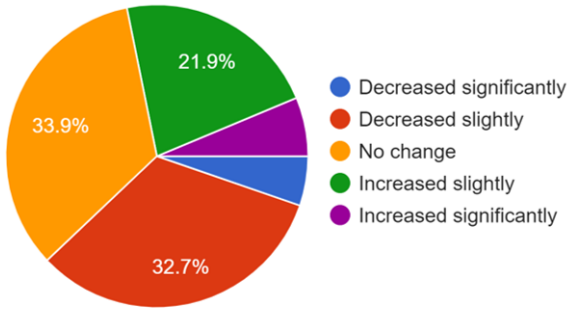

Fig. 3. Split across changes in Appetite

3) **criterion:** This parameter is used to specify the metric to be used while checking for the appropriate feature to split nodes in ensemble methods. Two possible values are gini ratio and entropy.
4) **max_depth:** This parameter is used to specify the maximum height of the decision tree beyond which splitting should not take place. This parameter is used to prevent overfitting in ensemble methods.
5) **max_features:** This is the number of features that are considered on a per-split level, rather than on the entire decision tree construction. During the construction of each decision tree, Random Forest will still use all the features (n_features), but it only consider number of "max_features" features for node splitting. And the "max_features" features are randomly selected from the entire features
6) **kernel:**Specific to support vector machines, this parameter specifies how the existing data must be transformed to a higher dimension.The Gaussian radial basis function (RBF) kernel is the most commonly used method in kernelized machine learning and SVM-specific problems.

Same exercise was repeated for depression and stress models as well and the k-fold accuracy for all the models is summarized in *Table 1*.

TABLE I
CLASSIFICATION ACCURACIES % (K-FOLD=10) FOR MODELS USED TO PREDICT DEPRESSION, ANXIETY AND STRESS LEVELS

Model	Anxiety	Depression	Stress
Logistic Regression	80.6	78.5	78.6
Support Vector Machine	79.6	78.8	78.5
Decision Tree	75.7	77.5	70.9
Random Forest	72.7	72.9	69.5

As evident from the results, the Logistic Regression model performs the best, giving the highest accuracy. Additionally, the logistic regression model offers better interpretation of the results as it allows the calculation of p-values and coefficients of each independent attribute, thus providing insights into the directionality and relevance of each independent attribute with respect to the target variable.

E. Population Segmentation

The treatment gap for any mental health concern in India was reported to be as high as 83% while 1% of the total population is at high suicide risk [14]. Therefore, it is crucial to identify the high-risk or high severity level groups so that proper help and resources could be provided which could help in the reduction of suicide rates, which in India was 15.7/100,000 (in 2015), higher than the global average of 10.6 [15]. This is further highlighted by the fact that the people going through stress/depression don't notice the symptoms and hence do not seek help (which is shown by the subject responses vs the DASS 21 survey)

For simplicity, the results and plots are only discussed for the Anxiety vertical, whereas the same approach is used for the depression and stress verticals. A segmentation approach is proposed using k-means clustering to identify different groups of varying severity. Elbow curve was used to identify the right number of clusters, which was found to be 10 as shown in *Figure 4.*.

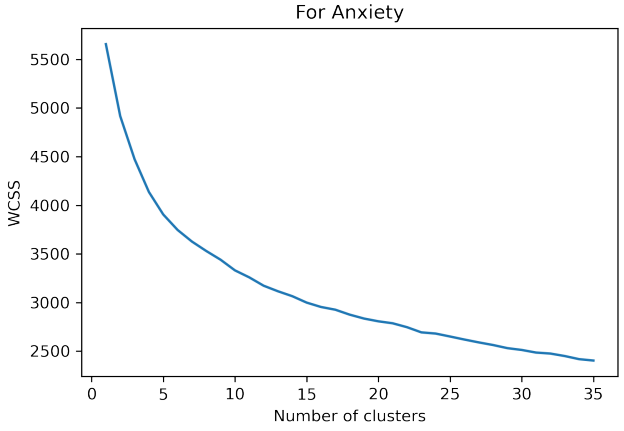

Fig. 4. Elbow Curve: WCSS vs number of clusters

III. Results and Discussion

The results of the logit model were analysed by means of Receiver Operating Characteristics (ROC) curves. In addition to the K-fold accuracy which was calculated, ROC-AUC curves give a good measure of the power of the model in terms of its ability to distinguish the various target classes from each other. ROC is a probability curve and in reference to our study, higher the area under the curve (AUC), the better the model is at distinguishing between respondents suffering from no anxiety from those with mild, moderate and severe anxiety, depression or stress. The ROC curve for Anxiety is shown in *Figure 5* and the AUC scores for all three models are given in *Table 2*. An AUC value above 0.9 is considered excellent for a predictive model [16].

The clustering results obtained through our model highlight various insights about the population that were unusual and don't go along with psychological trends. The clusters formed range across all parameters and assign different depression/stress/anxiety levels to segments of the population.

TABLE II
AUC scores of Logistic Regression on mental health metrics

Metric	AUC Score
Anxiety	0.94
Depression	0.93
Stress	0.92

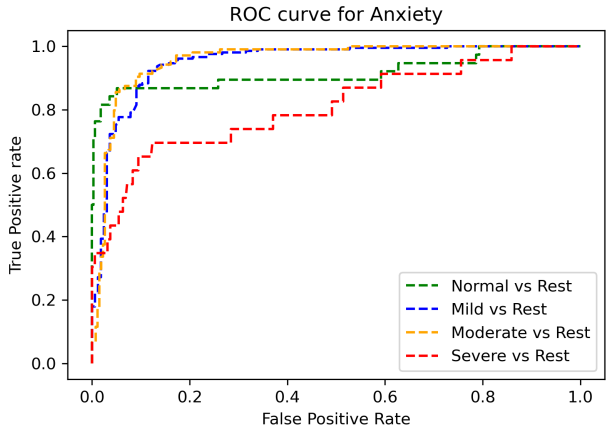

Fig. 5. ROC curve for different levels of anxiety

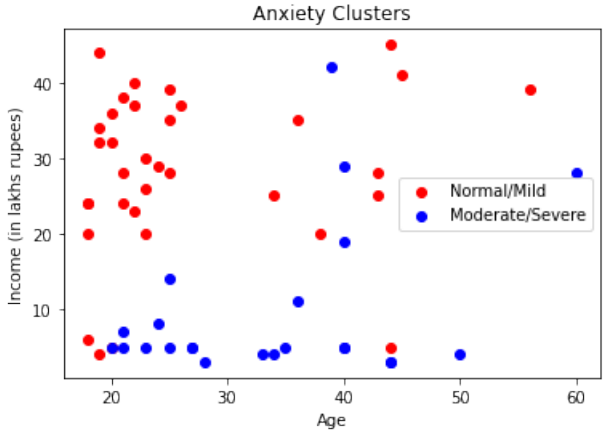

Fig. 6. Age vs. Income (simplified plot of clustering with only 2 clusters)

In *Figure 6*, the cluster group in blue with very low average annual income, when compared with the level of anxiety obtained from DASS-21 questionnaire revealed that the population in this cluster group had moderate/severe anxiety and similar exercise for the red cluster group, which had high average annual income, revealing that all the population in this cluster group had normal/mild anxiety.

Other dominant attributes of the clustering group in blue (i.e., group with severe/moderate anxiety) were: Age: 25-45 years old, Gender: Male, Occupation: Private, Income: less

TABLE III
AUC SCORES OF LOGISTIC REGRESSION ON MENTAL HEALTH METRICS

Anxiety	Depression	Stress
Directly proportional attributes		
Male	Age	Male
Private job	Private job	Unemployed
Student	Homemaker	-
Unemployed	Unemployed	-
	Self-employed	
Decrease in income during COVID 19		
Diabetes, BP or Heart issue	Diabetes, BP or Heart issue	Diabetes
Impact of COVID in life	-	Impact of COVID in life
Inversely proportional attributes		
Earning members	Annual household income	Earning members
Vaccination for COVID-19		
Doing exercise / yoga		
Appetite	-	Appetite
Sleeping hours	-	Sleeping hours

than 5 lakh rupees, Income change: Decreased significantly, no exercise or social interaction, reduced sleep and major impact of COVID-19 in life. This cluster of the Indian Population is the most prone to psychological stress.

Table 3 shows a list of the attributes which are statistically significant (p-value < 0.05) for severe level of anxiety, depression and stress. Directly proportional attributes are the attributes which increase the probability of falling into the severe category of anxiety/depression/stress, and inversely proportional attributes decrease the same probability. Following observations were made for severe anxiety:

- Males were more anxious than females. A study (Singh, 2020) also found that in India, men have a higher prevalence of mental illnesses than women [17]. Private sector employees, students and unemployed subjects had severe anxiety.
- The Decrease in income due to COVID-19 was also a driver of severe anxiety. People with prevailing health issues such as diabetes, blood pressure, heart problem etc. were found to be more prone to severe anxiety problems.
- Vaccinated people and people who do regular exercise/yoga had none to mild anxiety problems. Also, having adequate appetite and sleep helps in reducing anxiety.
- Unemployed people have very high depression and stress conditions due to various obvious factors.

As mentioned earlier, a total of 10 clusters were chosen to segment the population while the number of attributes were 15. It is very difficult to visualize the results of k-means clustering for this multidimensional data. Hence, *Figure 6*, shows a simple Age vs Annual Income plot for 2 cluster groups (out of 10) for anxiety. Some unusual insights that we observed from the clustering [18] applied is-

- Two of our clusters show that subjects with High Annual income and no prevailing health issues also experience severe depression (Cluster 9) and stress (Cluster 9) conditions. This could be due to various other factors which are not captured by the DASS-21 questionnaire but are prevalent in the population.
- One of the clusters (Cluster 7) show that the unemployed subjects have low anxiety conditions but have high stress and depression values. This can point to various socio-economic features such as the number of jobs available, and low acceptance rates.

It is seen that through Machine Learning, more insights from the data are inferred which are generally not captured by surveys, questionnaires and various other psychological measures. Due to deteriorating conditions, there should be plans in place to provide free help to the population. As it is shown from our study that, most of the people that do not know that they are under stress or depression are actually experiencing symptoms. The suggestion from this paper is that, from cluster to cluster, Private companies should have compulsory therapy sessions to reduce the distress caused by the pandemic.

IV. CONCLUSION

As it can be seen from the results of the logistic regression and K-means clustering pipeline, that the key attributes responsible for severe psychological conditions are identified. This research helps in disseminating information about the stigma that is created around Mental health in the Indian population. We identified, through clustering, the socio-economic factors of the high-risk to low-risk sections with a holistic analysis of the effect of the pandemic. In the current data-driven world, predictive models like the one suggested in this study could be beneficial in timely diagnosing serious mental health problems, especially for Indian population where people are still very reluctant and stigmatic about mental health issues.

These preliminary results can lead to a better interpretation of the mental health treatment needs of the people in India and could help healthcare professionals. private companies, government and policy makers to propose and implement new mental healthcare programs which could help the high-risk population by providing them better support and treatment options at subsidised rates. A future direction, in which the paper can develop to, is linking uncertainty in operations in which the majority of the Indian population is employed, to mental health [19].

REFERENCES

[1] R. S. et al., "India state-level disease burden initiative mental disorders collaborators. the burden of mental disorders across the states of india: the global burden of disease study 1990-2017," *Lancet Psychiatry*, vol. 7(2), no. 3, pp. 148–161, 2020.

[2] D. Nath, "Covid-19 — anxiety, depression top concerns on govt. helpline amid second wave, 2021," 2021.

[3] A. Gopal, A. J. Sharma, and M. A. Subramanyam, "Dynamics of psychological responses to covid-19 in india: A longitudinal study," *PLOS ONE*, vol. 15, no. 10, pp. 1–15, 10 2020.

[4] M. Schäfer S.K., Sopp, C. Schanz, M. Staginnus, A. Göritz, and T. Michael, "Impact of covid-19 on public mental health and the buffering effect of a sense of coherence," *Psychother Psychosom*, vol. 89, no. 6, pp. 386–392, 2020.

[5] P. Mahajan, P. Rajendran, B. Sunderamurthy, S. Keshavan, and J. Bazroy, "Analyzing Indian mental health systems: Reflecting, learning, and working towards a better future," *Journal of Current Research in Scientific Medicine*, vol. 5, no. 1, pp. 4–12, 2019.

[6] S. M. Gaiha, T. Taylor Salisbury, M. Koschorke, U. Raman, and M. Petticrew, "Stigma associated with mental health problems among young people in india: a systematic review of magnitude, manifestations and recommendations," *BMC Psychiatry*, vol. 20, no. 538, 2020.
[7] N. N. Author, "Covid-19 disrupting mental health services in most countries, who survey," 2020.
[8] J. P. . F. W. Chang Su, Zhenxing Xu, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, no. 116, pp. 2158–3188, 2020.
[9] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020, international Conference on Computational Intelligence and Data Science.
[10] K. Moghe, D. Kotecha, and M. Patil, "Covid-19 and mental health: A study of its impact on students in maharashtra, india," *medRxiv*, 2021.
[11] S. H. Lovibond and P. F. Lovibond, *Manual for the depression anxiety stress scales*, 2nd ed. Sydney, N.S.W. : Psychology Foundation of Australia, ©1995, 1995.
[12] A. G.-R. MM Rodríguez del Águila, "Sample size calculation," *Allergol Immunopathol (Madr)*, vol. 42(5), pp. 485–492, 2014.
[13] P. Khurana and D. Pandey, "Covid survey dataset public," 8 2021. [Online]. Available: https://github.com/Pranav2724/Covid-19-Research-Survey-Dataset
[14] O. P. Singh, "Closing treatment gap of mental disorders in india: Opportunity in new competency-based medical council of india curriculum," *Indian journal of psychiatry*, vol. 60, pp. 375–376, 2018.
[15] B. P. Srivastava K, Chatterjee K, "Mental health awareness: The indian scenario," *Industrial psychiatry journal*, vol. 25(2), p. 131–134, 2016.
[16] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010.
[17] U. A. Singh, "Disentangling india's mental health distress: Research on mental health: Is india catching up?" 2020.
[18] P. Khurana and D. Pandey, "Covid survey cluster analysis," 8 2021. [Online]. Available: https://github.com/Pranav2724/COVID-19-Research-cluster-analysis/
[19] K. Shariatmadar, A. Misra, F. Debrouwere, and M. Versteyhe, "Optimal modelling of process variations in industry 4.0 facility under advanced p-box uncertainty," in *2019 IEEE Student Conference on Research and Development (SCOReD)*, 2019, pp. 180–185.