Analysis of Mental Health during Lockdown

CIA 2

By

Adarsh Suresh 2048001
Tony Jose 2048021

Christ (Deemed to be University)
Central Campus, Bangalore

April 2021

CONTENTS

List of Figures

List of Tables

1. Introduction

The unprecedented pandemic of COVID-19 has caused schools and businesses closure and restriction on movement regarding containing this pandemic. Due to force lockdown, quarantine, loss of income, and regulation from contact with dying family members, people experience severe negative emotions, which increase the rate of mental stress. The pandemic and lockdown present a serious threat to the mental health of people, who may develop elevated rates of anxiety, depression, stress disorder, or even suicidal behaviors. Many studies have focused mainly on clinical and epidemiological research where mental health issues remain neglected.

Besides, there is a severe scarcity of information on the psychological impacts of the disease on population around the world. Wide varieties of individual and structural variables moderate this risk. In recent years, interest in artificial intelligence-aided health monitoring or psychological counseling systems has increased due to the convenience and efficiency of machine learning-based algorithms. Machine learning to mental health has demonstrated a range of benefits across the areas of diagnosis, treatment and support, research, and clinical administration. With the majority of studies identified focusing on the detection and diagnosis of mental health conditions, it is evident that there is significant room for the application of machine learning to other areas of psychology and mental health. In this cross-sectional study, we collected data through an online survey during the pandemic. Dataset was analyzed using machine learning algorithms aiming to detect stress which may affect one's mental health. We implemented a machine learning approach regarding factors associated with different mental health outcomes in a pandemic situation.

1.1. Problem Statement:

COVID-19 is imposing a threat both on physical and mental health since its outbreak. The world has adopted a lockdown strategy with potential consequences on day-to-day life, mental and physical health. This survey aims to explore the impact of lockdown on mental health among people.

1.2. Need for the study:

Currently, there are no reports examining the psychological impact and response of people during the peak of the COVID-19 epidemic. This survey provides information about the risk factors associated with psychological impacts that should be helpful for concerned authorities to plan and adopt appropriate interventions to overcome the negative psychological impacts to ensure sound mental health.

2. About the data

Using a Google form, we prepared a standard questionnaire and the link was distributed on social media platforms and by texting on private networks. Our questionnaire included 30 questions on various aspects, self-assessment of anxiety and stress, the impact of COVID and lockdown on their lifestyle, related physical stress, the attitude towards suicide. This form was aimed to collect data on the concerns of people about the effects of lockdown on their mental well-being and the possible ways to maintain the mental well-being of people. The goal is to predict whether future patients have mental issues by extracting information from the relationship between various stress levels in our dataset. Information was collected from nationwide citizen, irrespective of the state (province) of residence, caste, creed, religion, and sex; the researchers did not consciously exclude any social group. The respondents represented urban, semi urban and rural areas. It required an estimated mean time of 3 to 5 minutes to fill out and submit. We received complete responses from 471 participants (54.2% female and 45.4 male). Table 2.1 and Table 2.2 given below shows the dataset description and feature description respectively.

Table. 2.1. Dataset description

SI No	Criteria	Details
1	Name	Analysis of Mental Health during Lockdown
2	Type	Multivariate
3	No of rows	471
4	No of columns	21
5	Missing values	NILL
6	Target type	Categorical
7	Applicable technique	Classification

Table. 2.2. Feature description

Sl No.	Feature Name	Description
1	Name	Indicates name of the participants
2	Age	Indicates age group of the participants. Having categories Below 18, 18-25, 26-32, 33-39, 40-50 and above 50.
3	Gender	It takes 3 unique values, Male, Female and Other
4	State	Indicates geographic location of the participants
5	Where do you live	Indicates the region of living
6	Marital Status	Indicates the participant married or not
7	Occupation	Indicates the profession of the participants
8	Time spend on electronic devices	Amount of time spend on screen
9	How often do you feel lonely	Specifying the frequency of the participant feeling lonely
10	How often do you feel stressed?	Target variable indicating the stress level of the participants
11	Does your day-to-day work routine affect your stress level?	Shows the relation between work and stress
12	Where do you think the stress levels are experienced more?	Specifying the place where stress is experienced more
13	Do you lack energy and motivation?	Enquiry about motivation and energy level of the participant
14	Have you had thoughts about harming yourself after being stressed out?	Shows suicidal tendency of the participants
15	Do you have trouble getting to sleep and staying asleep?	Indicates the trouble in sleep
16	What are your coping strategies to overcome this stress?	Measures to overcome stress, taken by participants
17	Were you affected by covid19?	Indicates whether the participant were affected by Covi or not
18	Did you face any physical health issues other than covid19 during lockdown? If yes please specify else mention no	Information about health issue other than Covid19
19	Did you experience some or all of the following?	Enquiry about their mental state
20	What about managing finance during lockdown?	Financial crisis faced by the participant

3. Implementation details

We got a set of complete responses from 471 people. Data were analyzed aiming to gain information concerning serious stress leading indications and their potential relationships with mental health. We tested machine learning (ML) models to predict the stress among people during the pandemic (lockdown) and to evaluate our proposed model. We have correlated stress with different attributes. In this analysis, we looked at more stimulating data and progressive techniques and interpretations. It involved building the model, testing the model (performance evaluation), and comparing it with other machine learning models by accuracy score.

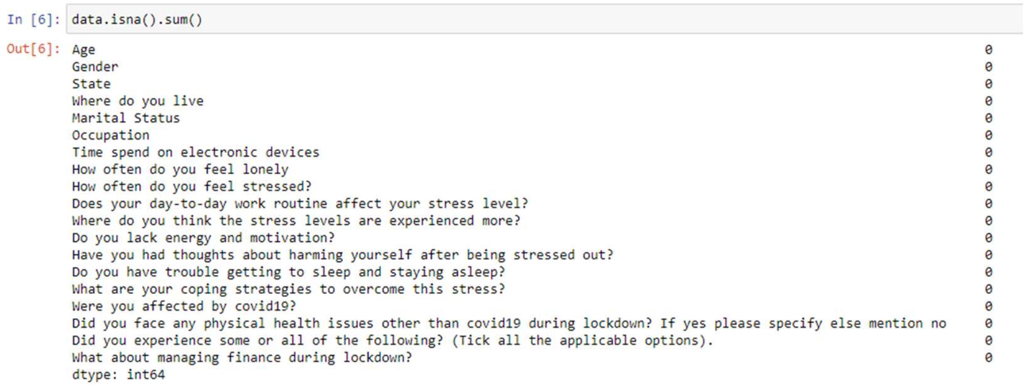
All analyses were accomplished in Python language on the Jupyter Platform. Packages used are **matplotlib, NumPy, seaborn, sci-kit learn, pandas**. First, the data explored for any null values as well as datasets were cleaned. The data was not equally distributed therefore different methods of sampling were made. The variable to be predicted was categorical and hence different algorithms such as **Random Forest, Logistic Regression, KNN and Support Vector Machines (SVM)** were used for modelling the data. Data split into a training set (70%), and a test-set (30%). The data frame which has both the independent and dependent variables were splitted to X_train, y_train, X_test and y_test. Each model was trained by inputting X_train, y_train values of the training dataset. Model evaluation and the confusion matrix were drawn for each algorithm. For boosting the accuracy, a model using **XGBoost** is also made and fitted for the test set.

3.1. Preprocessing

This is a non-prepared data set so necessary cleaning should be done for better analysis. As this is a classification problem first we identified the unwanted variables like Name, State etc. and filtered out from the data set as an initial cleaning.

3.1.1. Missing values

One of the important steps in data preprocessing is to find out and handle the missing data in the data set as these values can create false interpretations. The column wise missing value details are given below in Figure 3.1. As a result, there were no missing values found in the data set.



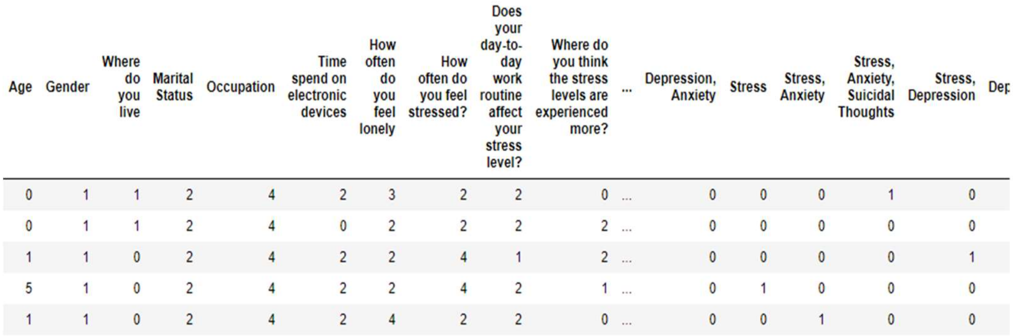
```
In [6]: data.isna().sum()

Out[6]: Age                                                                                      0
        Gender                                                                                   0
        State                                                                                    0
        Where do you live                                                                        0
        Marital Status                                                                           0
        Occupation                                                                               0
        Time spend on electronic devices                                                         0
        How often do you feel lonely                                                             0
        How often do you feel stressed?                                                          0
        Does your day-to-day work routine affect your stress level?                              0
        Where do you think the stress levels are experienced more?                               0
        Do you lack energy and motivation?                                                       0
        Have you had thoughts about harming yourself after being stressed out?                   0
        Do you have trouble getting to sleep and staying asleep?                                 0
        What are your coping strategies to overcome this stress?                                 0
        Were you affected by covid19?                                                            0
        Did you face any physical health issues other than covid19 during lockdown? If yes please specify else mention no    0
        Did you experience some or all of the following? (Tick all the applicable options).      0
        What about managing finance during lockdown?                                             0
        dtype: int64
```

Fig. 3.1 Column wise missing value details

3.1.2. Label encoding

The categorical variables of both the data frames have been converted to numerical by Label Encoding. Figure 3.2 shows the cleaned data.

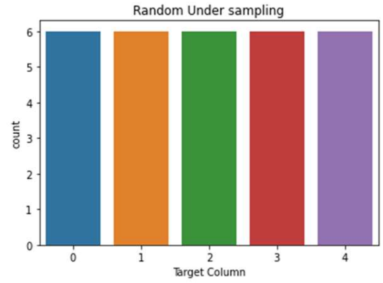


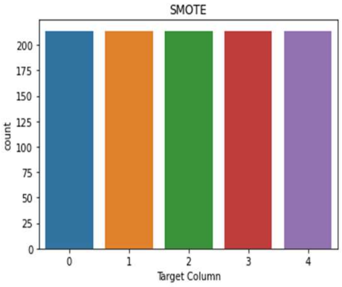
Age	Gender	Where do you live	Marital Status	Occupation	Time spend on electronic devices	How often do you feel lonely	How often do you feel stressed?	Does your day-to-day work routine affect your stress level?	Where do you think the stress levels are experienced more?	...	Depression, Anxiety	Stress	Stress, Anxiety	Stress, Anxiety, Suicidal Thoughts	Stress, Depression	Dep
0	1	1	2	4	2	3	2	2	0	...	0	0	0	1	0	
0	1	1	2	4	0	2	2	2	2	...	0	0	0	0	0	
1	1	0	2	4	2	2	4	1	2	...	0	0	0	0	1	
5	1	0	2	4	2	2	4	2	1	...	0	1	0	0	0	
1	1	0	2	4	2	4	2	2	0	...	0	0	1	0	0	

Fig. 3.2 Cleaned data

3.1.3. Sampling data

Due to irregular distribution of samples, we have tried three different sampling i.e., Random under-sampling, SMOTE and Random Over-sampling. Figure 3.3, 3.4 and 3.5 shows the distribution of the target column after applying the sampling technique.





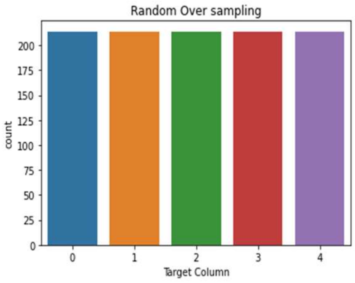


Fig. 3.3 Random Under Sampling Fig. 3.4 SMOTE Fig. 3.5 Random Over Sampling

The best sample was given by Random Over-Sampling. So we use the Random Over-Sampling model for our model building.

3.2. EDA

Exploratory data analysis is a statistical operation to outline tendencies within data, assisted by visualizations. EDA was essentially applied on the dataset for extracting insights beyond formal modeling to hypothesize features reinforced by data.

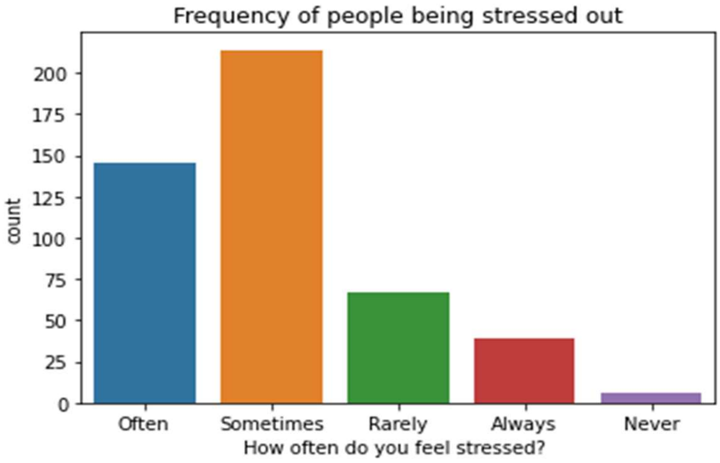


Fig. 3.6 Frequency of people being stressed out

Figure 3.6 shows that a large number of people were stressed during the COVID19 lockdown time. Only very few people were reported, of never feeling stressed.

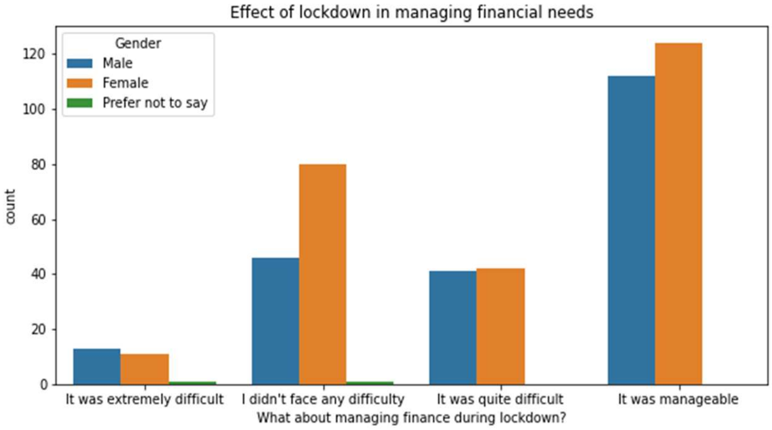


Fig. 3.7 Effect of lockdown in managing financial needs

From Figure 3.7 we can see that most of the people who participated in the survey didn't face much of a financial crisis. People who have faced financial crises come less than 25% of the total population.

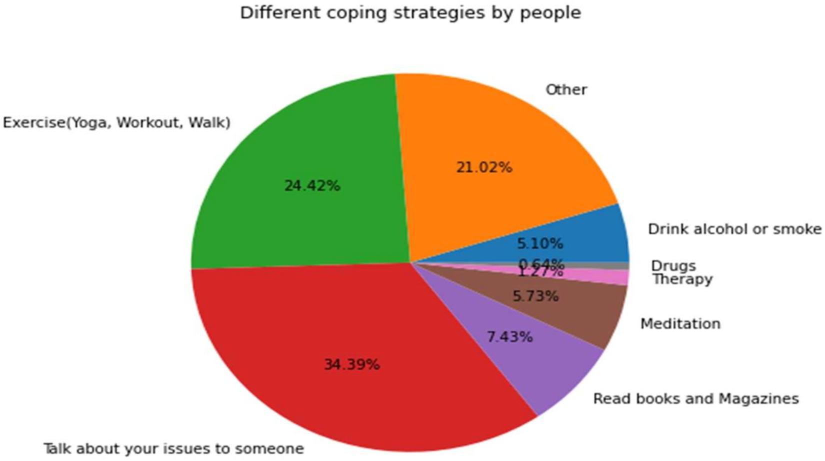


Fig. 3.8 Different coping strategies by people

The above Figure 3.8 shows different measures or ways in which people deal with Stress. More than 34% of people depend on “Talking to someone about their issues” more effective in dealing with stress. About 24% depends Exercise (Yoga, Workout, and Walk) as a way to get rid of stress. Only about 2% of the population seek medical assistance.

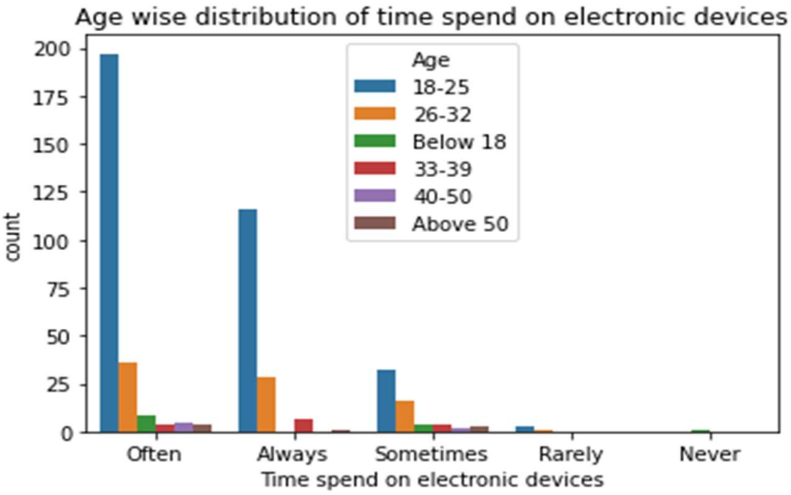


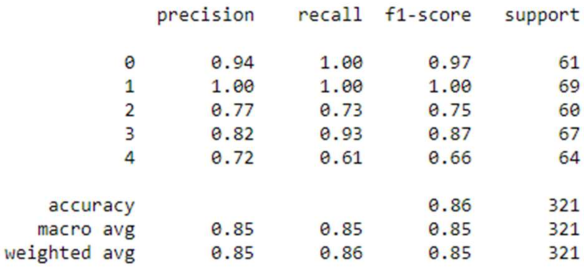
Fig. 3.9 Age wise distribution of time spend on electronic devices

Figure 3.9 indicates that people in the age group of 18-25 have more screen time when compared with other age groups. It is very clear that people aged more than 50 have a very less tendency to use electronic devices.

3.3. Model building and evaluation

We built the supervised Machine learning model which can predict mental health in advance. For the same we divided the data into training and testing in a 70-30 ratio, and removed the class imbalance in the training set, we performed over sampling using the Random Over-Sampling algorithm in the dataset. Using the training dataset, we train our model. And further, we evaluated the performance of the model in the test dataset.

Fig. 3.10 given below, shows the classification report when the model was made with the help of Random Forest Classifier. It gave an accuracy of almost 86%.



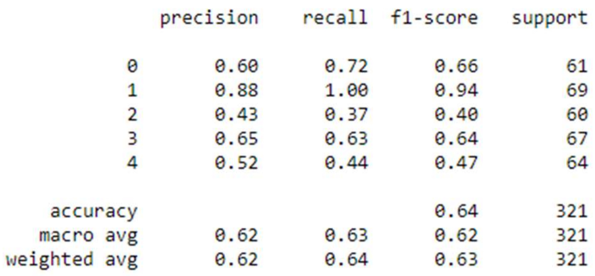
```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        61
           1       1.00      1.00      1.00        69
           2       0.77      0.73      0.75        60
           3       0.82      0.93      0.87        67
           4       0.72      0.61      0.66        64

    accuracy                           0.86       321
   macro avg       0.85      0.85      0.85       321
weighted avg       0.85      0.86      0.85       321
```

Fig. 3.10 Classification table of Random Forest Classifier

Fig. 3.11 given below, shows the classification report when the model was made with the help of Logistic Regression. It gave an accuracy of almost 64%.



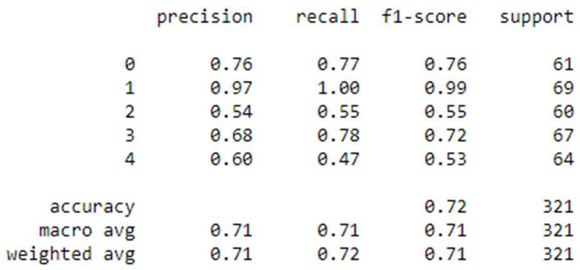
```
              precision    recall  f1-score   support

           0       0.60      0.72      0.66        61
           1       0.88      1.00      0.94        69
           2       0.43      0.37      0.40        60
           3       0.65      0.63      0.64        67
           4       0.52      0.44      0.47        64

    accuracy                           0.64       321
   macro avg       0.62      0.63      0.62       321
weighted avg       0.62      0.64      0.63       321
```

Fig. 3.11 Classification table of Logistic Regression

Fig. 3.12 given below, shows the classification report when the model was made with the help of Support Vector Classifier. It gave an accuracy of almost 72%.



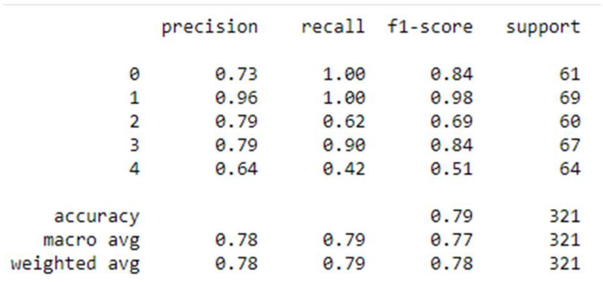
```
              precision    recall  f1-score   support

           0       0.76      0.77      0.76        61
           1       0.97      1.00      0.99        69
           2       0.54      0.55      0.55        60
           3       0.68      0.78      0.72        67
           4       0.60      0.47      0.53        64

    accuracy                           0.72       321
   macro avg       0.71      0.71      0.71       321
weighted avg       0.71      0.72      0.71       321
```

Fig. 3.12 Classification table of Support Vector Machine

Fig. 3.13 given below, shows the classification report when the model was made with the help of K-Nearest Neighbor Classifier. It gave an accuracy of almost 79%.



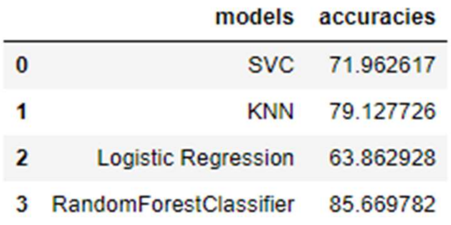
```
              precision    recall  f1-score   support

           0       0.73      1.00      0.84        61
           1       0.96      1.00      0.98        69
           2       0.79      0.62      0.69        60
           3       0.79      0.90      0.84        67
           4       0.64      0.42      0.51        64

    accuracy                           0.79       321
   macro avg       0.78      0.79      0.77       321
weighted avg       0.78      0.79      0.78       321
```

Fig. 3.13 Classification table of K-Nearest Neighbor

3.4. Comparative study



	models	accuracies
0	SVC	71.962617
1	KNN	79.127726
2	Logistic Regression	63.862928
3	RandomForestClassifier	85.669782

Fig. 3.14 Accuracies of models made

The accuracies acquired by different classification models is given in Figure 3.14. After analyzing the models, we can see that Random Forest Classifier has given us more accurate results than the other three models. Not only in terms of accuracy but also the time complexity was less when compared with other models. It can handle binary features, categorical features, and numerical features. There is very little pre-processing that needs to be done. The data does not need to be rescaled or transformed. Second highest accuracy was given by the KNN classifier followed by the SVC. The main drawback of SVC and KNN was that the time taken by model to fit the data is very high. The least accurate model was given by Logistic regression. It might be because the model attempts to predict precise probabilistic outcomes based on independent features. All these results indicate a reasonable goodness-of-fit for the Random Forest model and thus it performs better than the other three models.

4. Results and Discussion

This study detected the stress based on the collected data and analyzed factors that lead to excessive stress and affecting mental health. We detected the stress with the help of machine learning model. Classification model and is built, tested, in the interest of diagnosing stress. The performance is tested and compared with other models. Our model achieved its best performance in detecting stress.

The Table 4.1 given below shows the final results after performing the analysis on the dataset.

Table. 4.1 Final result of different models

Sampling Technique	Mode			
	Random Forest	Logistic Regression	SVC	KNN
Random Under-Sampling	33.33	22.22	33.33	22.22
Random Over-Sampling	85.66	63.86	71.69	79.12
SMOTE	81.3	69.47	70.09	73.2

After performing all the preprocessing steps, due to the imbalance in the data we tried three different sampling techniques to balance the data. Using the technique of Random Under-Sampling, the results weren't very satisfying. So we tried sampling using methods of Random Over-Sampling and SMOTE. The sampled values were the same in both the sampling techniques. To finalize which technique to be used we followed trial and error method. So, four models were built for both the sampling techniques and the one which gave better accuracy was selected. The results were almost similar but then better and accurate results were given by Random Over-Sampling technique. Using this method, the accuracies of Random Forest Classifier, Logistic regression, Support Vector Classifier and K-Nearest Neighbor were 85.66%, 63.86%, 71.69% and 79.12% respectively.

5. Limitations

The present study has several limitations. First, the number of people who participated in the survey was small, and the possible reason is the voluntary nature of the survey. Hence, some people opted not to participate. Furthermore, the voluntary nature of the survey also made the follow-ups difficult to conduct (very few people were willing to fill in the questionnaire again). The participants were recruited via an online link posted on social networks and other social media platforms. While online recruitment guarantees large samples, it does not guarantee sample representativeness. For this reason, very vulnerable groups, such as people in the remote areas and people who do not have technological access, may not be well represented in this study. The use of self-report measures did not enable us to verify the reliability of the responses, or to ensure that participants correctly understood the questions. And finally, due to the fact that this survey was related to mental health and stress, many people refused to take part in the survey which led to a decrease in the number of responses.

6. Conclusions

In this research, four machine learning models, namely, Random Forest, K-Nearest Neighbor, Logistic Regression and Support Vector classifier, were systematically analyzed and compared for classifying mental stress during Covid19 lockdown. According to this case study, all four models exhibit reasonably good performances; the Random Forest model has the highest accuracy compared with the other three models. The Random Forest model, with an accuracy score of 85.66%, is a promising technique for classifying the stress rate among populations. Finally, these study results may be useful for decision makers and health officials in helping people overcome or in early detection of mental stress.

7. Future works

- Future researchers can test this model using other supervised learning models as well as unsupervised learning models.
- Accuracies can be further increased by the usage of boosting algorithms such as Adaboost, XGBoost etc.
- With the help of some professionals, we can add more relevant features and can further update the model.

References

[1] https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/#:~:text=Balance%20data%20with%20the%20imbalanced,thus%20seeking%20to%20preserve%20information.

[2] Rajkumar RP. COVID-19 and mental health: A review of the existing literature. Asian J Psychiatr 2020;52:102066. https://doi.org/10.1016/j.ajp.2020.102066.

[3] Gupta M, Vaikole S. Recognition of Human Mental Stress Using Machine Learning Paradigms. SSRN Electron J 2020. https://doi.org/10.2139/ssrn.3571754.

[4] Chao Chen AL and LB. Using Random Forest to Learn Imbalanced Data | Department of Statistics 2004. https://statistics.berkeley.edu/tech-reports/666 (accessed January 7, 2021).

[5] Galea S, Merchant RM, Lurie N. how-machine-learning-changing-mental-health conditions. JAMA Intern Med 2020;180:817–8. https://doi.org/10.1001/jamainternmed.2020.1562.

[6] Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artif Intell Med 2018;90:1–14. https://doi.org/10.1016/j.artmed.2018.06.002.