Comparative Analysis of Stress among Undergraduate Students Using Logistic Regression and Random Forest Techniques

Uzma Uzma, Subhabrata Kanjilal, Nishi Yadav*

[1]School of Studies in Engineering and Technology, Guru Ghasidas Vishwavidyalaya Bilaspur, India

uzma.shine2017@gmail.com, rohan525kanjilal@gmail.com, nyadav.cse@gmail.com

Abstract: Mental illness has become a major problem for youngsters nowadays. Our project deals with calculation of stress as we know that overall collegiate performance and social obligation have created a pressurized cerebral as well as emotional state for students. With limited college seats, and high number of post metric students applying to get into the top universities and colleges, it could be difficult to get into the college one wished for. Same is the case for a student in his last year of graduation. There is lot of pressure that one undergoes like pressure of getting placed, pressure of getting into a top college for PG and many more. Lastly the stress caused due to the pandemic can be least ignored. Students weren't able to attend online classes properly due to lack of resources which resulted students to undergo a lot of stress about their academics. We collected the data through a google survey form which was send to all the students known to us. We collected 101 responses and then converted them into numerical value and lastly implemented the logistic regression and random forest algorithm, where we got our f1 score = 0.9411 and Accuracy score (Random Forest) = .0.8571.

Keywords— Stress, Logistic regression, Random Forest, Mental illness, Predictive Analysis, Stress in Students, Machine Learning, Data Analysis, Causes & Effects of Stress.

1. INTRODUCTION

Stress can be described as a condition where someone is mentally ill because of the adversities/changes that he/she is undergoing or underwent [1]. There are many types of stress such as physical stress which includes trauma, endocrine and/or biochemical disparity, nutritive stress, Exsiccosis, chemical abuse, toothache, and human locomotor system, misalignment or imbalance. The second type can be psychological stress which includes emotional stress, cognitive stress, and perceptual stress. Third category can be Psychosocial stress: Relationship/marriage difficulties, lacking of social help, lacking of assets for better survival, loss of job/money/saving, loss of your dearest, bankruptcy, house expropriate, and separation. Lastly the psycho-spiritual stress which can be a exigency of manners, significance, and motive; melancholy fighting (apart from being fruitful, rewarding, meaningful be gloomy); and also moving out of track within one's core faith or beliefs. As a optimistic result, stress can urge us to perform an operation that we wanted to do. As a negative impact it can lead to certain health related issues like headache, increase in heart rate, increase in blood pressure etc. Due to these health issue sometimes, the person consecutively leads to depression, anger or sometimes distrust. It differs from person to person so as how he/she deals with a particular situation. When a person gets too much happy or sad the stress within him/her readjusts itself in accordance with the current circumstances that the person is undergoing.

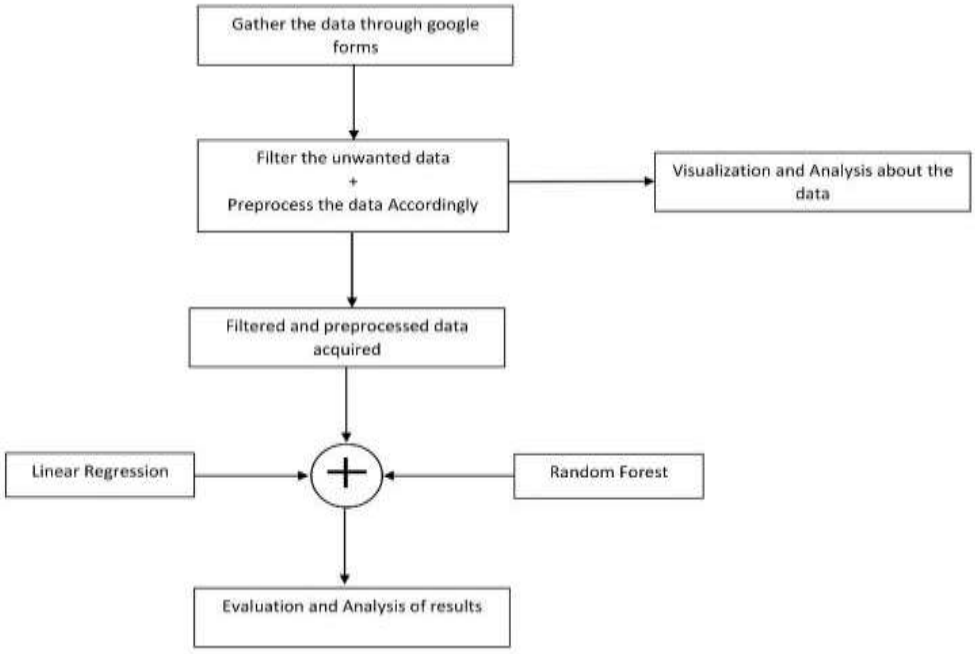


Fig. 1. Graphical Abstract

Table.1. Causes & Effects of Stress reported by students

Behavioral effects: - i. Alteration in activity levels ii. Decrease in ability and efficacy iii. Difficultly in communicating iv. Increased sense of humour/gallows humour v. Irritability, outbursts of anger, frequent arguments vi. Ability to nap, chill or let down vii. Modification in eating regime viii. Insomnia ix. Change in work execution x. Periods of bawl xi. Increment in intoxication, sugar or caffeine xii. Attentive about safety or the surrounding habitat xiii. Avoidance of things or places that prompt memories xiv. Accident prone	**Causes: -** i. Study overload ii. Lack of financial Support iii. Family issues iv. Issues with friend or significant partner v. Health Related Issue vi. Involvement in clubs/organisation
Psychological or emotional effects: - i. Feeling fearless, joyful or invincible ii. Contradiction in thoughts iii. Anxiety or alarmed iv. Worry about security of yourself and others v. Get irritated or angry easily vi. Anxious or nervous vii. Misery, sullen, regret or depression viii. Realistic or anguish dreams ix. Apologetic or "survivor guilt" x. Feeling speechless, incapable or desperate	**Cognitive effects: -** i. Memory problems/forgetfulness ii. Disorientation iii. Confusion iv. Slowness in thinking, analysing, or comprehending v. Difficulty in calculating, setting preference giving opinion vi. Difficulty focusing vii. Short notice span viii. Loss of aim

xi. Feeling deserted, off-track, desolate or abandoned **xii.** Boredom **xiii.** Compulsive behaviour Feeling misinterpreted or unacknowledged	ix. Foggy about the disaster or an incident happened
Bodily effects – i. Tachycardia ii. Increased blood pressure iii. Upset stomach, nausea, diarrhoea iv. Increased or decreased appetite which may be accompanied by weight loss or gain v. Sweating or chill vi. Tremors or muscle twitching vii. Muffled hearing viii. Tunnel vision ix. Feeling uncoordinated x. Headaches xi. Pain in muscles xii. Photosensitive sight xiii. Lumbago xiv. Globus sensation xv. Easily startled xvi. Tiredness that does not go with sleep xvii. Menstrual cycle changes Change In sexual desire or response	

2. MATERIALS AND METHODS

A. Study Population

We created a google form having 12 questions asking about the person's experience about how he dealt with the stress. We also tried to put certain questions where we enquired about whether he feels or not that he is in stress by scaling himself from the number 1 to 10 , where 1 meant not at all stressed and 10 meant very stressed. Similarly, we enquired about how much he can handle the stress by scaling himself from the number 1 to 10 , where 1 meant cannot handle at all and 10 meant can to handled easily . The questionnaire also dealt with what were the causes of stress and how did it affected the person

B. Filtering the data

We received about 109 responses but after filtering we were left with only 101. We went through each and every minute details of the responses received and deleted the responses which were duplicate or were not felt genuine. We then created another xml sheet which dealt with only the information required excluding the information like gender, study, year of birth etc.

C. Conversion of the data

The responses were then converted into numerical like if a person has ticked one option for the usual cause of stress as study issues, then we gave 1 to the vertical corresponding study issue and gave 0 to all other usual causes of stress. Similarly, it was done for all the left 81 responses. Finally, we added all the values of a particular row and stored them in their corresponding row. Then we took the average of all the values and concluded that a person whose values were added below 34 were not in stress whereas above 34 could be in stress. For the people who got a score below 34 we checked their answer for the question - How do you usually experience stress, if the answer to this particular was found relevant then too we concluded that the person is in stress even though he has a score below 34. As we can see in the screenshot below that

the first response was that the usual cause of stress for that student was issue with significant other partner so we gave 1 to this particular vertical and the left other columns were marked 0.

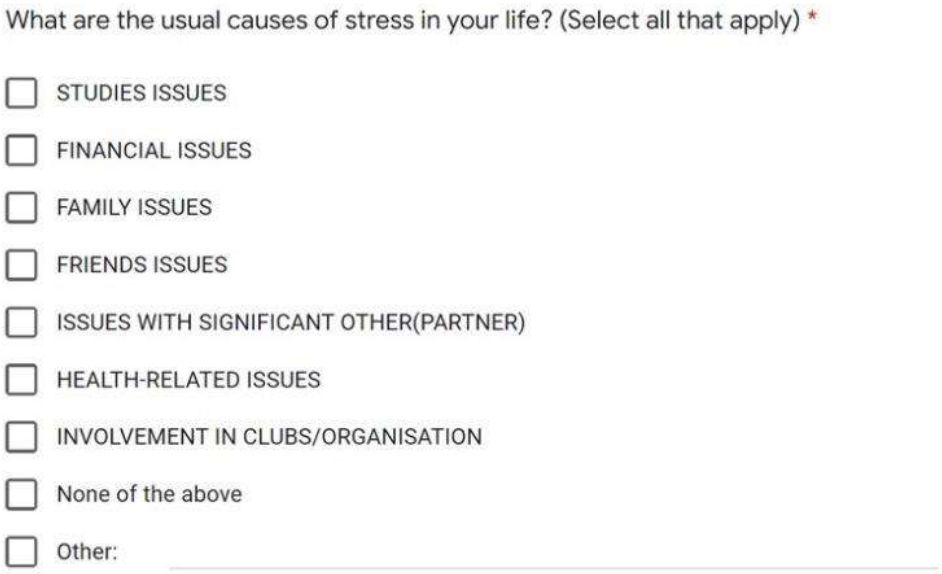
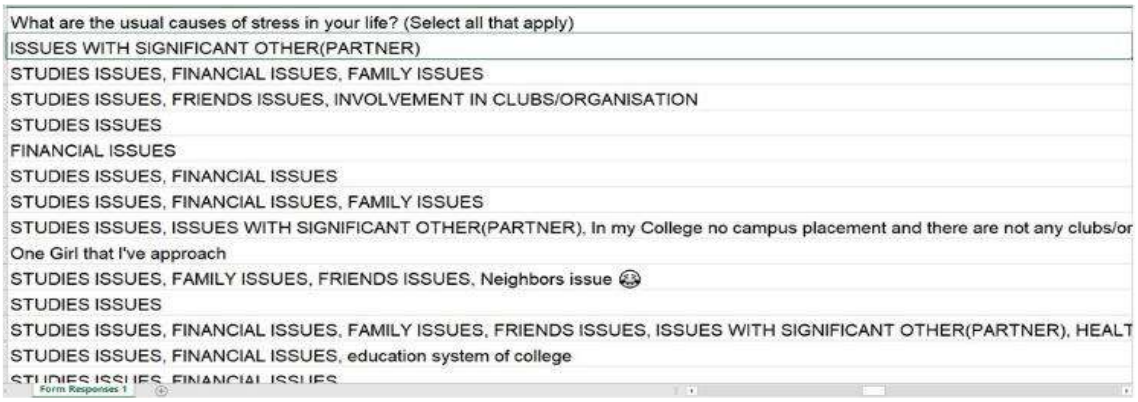
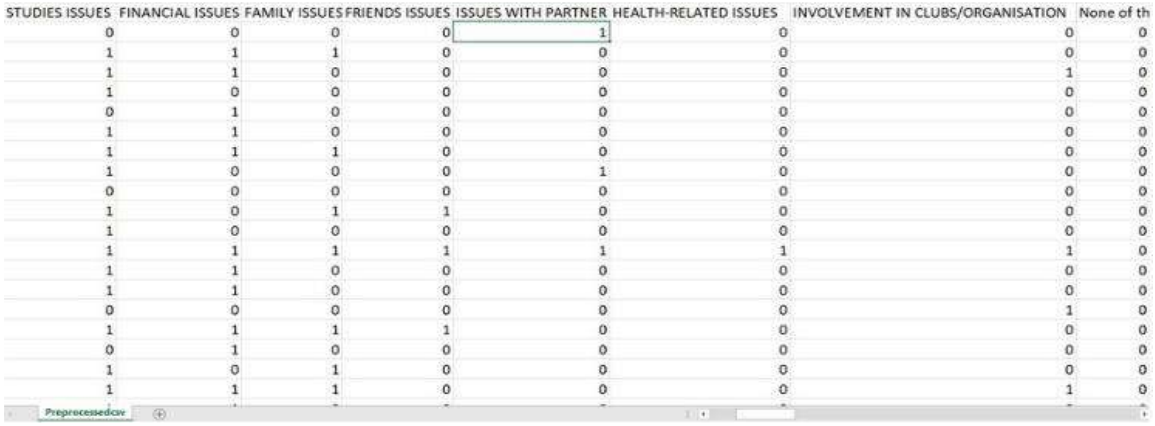


Fig. 2. Screenshot of the Google Form



What are the usual causes of stress in your life? (Select all that apply)
ISSUES WITH SIGNIFICANT OTHER(PARTNER)
STUDIES ISSUES, FINANCIAL ISSUES, FAMILY ISSUES
STUDIES ISSUES, FRIENDS ISSUES, INVOLVEMENT IN CLUBS/ORGANISATION
STUDIES ISSUES
FINANCIAL ISSUES
STUDIES ISSUES, FINANCIAL ISSUES
STUDIES ISSUES, FINANCIAL ISSUES, FAMILY ISSUES
STUDIES ISSUES, ISSUES WITH SIGNIFICANT OTHER(PARTNER), In my College no campus placement and there are not any clubs/or
One Girl that I've approach
STUDIES ISSUES, FAMILY ISSUES, FRIENDS ISSUES, Neighbors issue 😂
STUDIES ISSUES
STUDIES ISSUES, FINANCIAL ISSUES, FAMILY ISSUES, FRIENDS ISSUES, ISSUES WITH SIGNIFICANT OTHER(PARTNER), HEALT
STUDIES ISSUES, FINANCIAL ISSUES, education system of college
STUDIES ISSUES, FINANCIAL ISSUES

Fig. 3. Screenshot of the Responses



STUDIES ISSUES	FINANCIAL ISSUES	FAMILY ISSUES	FRIENDS ISSUES	ISSUES WITH PARTNER	HEALTH-RELATED ISSUES	INVOLVEMENT IN CLUBS/ORGANISATION	None of th
0	0	0	0	1	0	0	0
1	1	1	0	0	0	0	0
1	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0
1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0
1	0	0	0	0	0	0	0
1	1	1	1	1	1	1	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0
1	1	1	1	0	0	0	0
0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0
1	1	1	0	0	0	1	0

Fig. 4. Screenshot of the converted data

D. Related work

In [1], the author founded analytically notable difference between bottom line and stress periods for seven from the complete parameters used. Given the determined noteworthy similarities between some of the variables. They concluded that three EDA variables (average

SCL, number of 14 peaks and duration) and one HRV parameter (average HR/ average RR/HF) can be used in a hand-on stress analysis procedure [1].
In [2], the authors found that out of these four algorithms Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest. SVM has performed good as its quantity of data is and its geometric way of categorization is also low. Finding and Analyzing methods like PSS with increased exact conclusions and small cost can help improve the cerebral health of each person and make everyone sound fine (mentally) [2]. In [3], the authors developed a tool using Naive Bayes and Sentimental Analysis which was capable of classifying student tweets into a different set of categories based on the student expressed emotions [3]. In [4], the authors compared two algorithms that were Naive Bayes and K-Nearest Neighbors, after all the [13] processing done they concluded that the switch in the amount of information has affected accuracy, f1-score, recall & precision, both through the percentage split tests and k-cross validation [4]. In [5], the authors worked on the Logistic Regression, SVM (Linear Kernel), KNN, Decision Tree, Random Forest and Proposed Ensemble Model and they concluded that more than 74% of the students experienced stress [5]. In [6], the authors described stress as the actuality that is worldwide existed, among all the students of almost each and every stream, irrespective of gender, age, class and other attributes. They performed a parallelepiped questionnaire-based look-over before coming to this conclusion [6]. In [7], the authors used two methods namely the Coping Inventory for Stressful Situations (CISS) & Alcohol Use Dependency Identification Test (AUDIT), and concluded that there is no co-relation among drinking intentions, alcohol dependency risk and stress handling styles in the observed group of people [7]. In [8], the authors observed that the relationship between ACEs, and stress report by the individual was managed by the PTSD-S. This shows that students who report PTSD-S following childhood misfortune undergo higher levels of stress. Traditional and non-traditional students have a difference in their ICLRE scale responses [8].

3. DATA PROCESSING

A. Algorithms

Logistic regression has been used to estimate the probability that whether a student is suffering from high stress (1) or is in low stress (0), using the causes and effects that are thought to be related to or influence such cause [14]. Here in this paper, we have implemented Binary Logistic Regression which has a dependent variable “Stress” represented by a target variable, which has two values labelled ”0” which represents low stress and ”1” which represents high stress [9]. Logistic regression has been used to estimate the probability that whether a student is suffering from high stress (1) or is in low stress (0), using the causes and effects that are thought to be related to or influence such cause [14]. The second Algorithm used is Random Forest, it considers so many decision tresses thus forming a forest. It uses selecting feature randomly for building every individual decision tree and then try to create an uncorrelated forest of trees whose prediction by a certain group of features is more accurate than that of any individual decision tree [13].

B. Data Analysis

Stress	0	1
Gender		
Female	21	17
Male	35	28

Fig. 5. In the above crosstab, shows the no. of males and females under high stress (1) / low stress (0).

Stress	0	1
Age		
19	5	0
20	11	2
21	13	12
22	11	11
23	9	14
24	6	3
25	1	2
27	0	1

Fig.6. Age to Stress Comparison

In the above crosstab, we can conclude that people who are in age range of 18 -21, most of them are under low stress (0), the people who are in age range of 20-22, some of them are under low stress (0) and some of them are under high stress (1). But in the age range of 22-26, most of the people are in high stress (1).

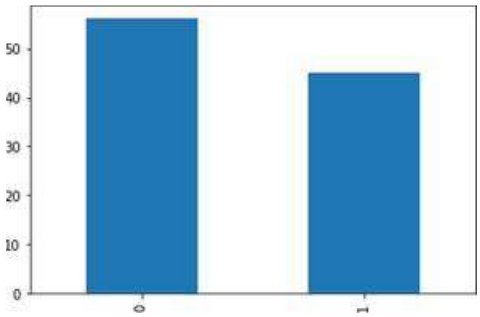


Fig. 7. Distribution of Stress

C. Performance Parameters

Accuracy: Overall, how often has the classifier correctly classified the data?

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \text{ [15] ---------- (1)}$$

where, TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative [15].

True Positive Rate (TPR): It can be defined as when the actually *data* denotes that the person is actually in low stress but the system predicts low stress?

$$TPR = \frac{TP}{TP+FN} \text{ [15] ---------- (2)}$$

Also known as" **Recall**" or" **Sensitivity**". [15]

False Positive Rate (FPR): It can be defined as when the actually data denotes that the person is actually in high stress but the system predicts low stress [15].

$$FPR = \frac{FP}{FP+TN} \text{ [15] ---------- (3)}$$

True Negative Rate(TNR): It can be defined as when the actually data denotes that the person is actually in high stress but the system predicts high stress.

$$TNR = \frac{TN}{FP+TN} \text{ [15] ---------- (4)}$$

It is equivalent to (1-FPR), also known as" **Specificity**" [15].

Precision: How much the data is actually low stressed, when the model predicts low stress correctly?

$$Precision = \frac{TP}{TP+FP} \text{ [15] ---------- (5)}$$

Precision will be calculated out of all the positive classes (i.e. low stress) the model predicted correctly, versus the number of classes that are actually positive (i.e. the data is actually of low stress) and Accuracy will be the number of class that were predicted correctly out of all the classes [16].

f1 Score: It is highly unfavorable to validate a model which a high recall and a low precision value or vice-versa. In order to make the model comparable, we use f1-Score. It is basically the harmonic mean between precision and recall [17] i.e.

$$f1\ score = \frac{2*Precision*Recall}{Precision+Recall} \text{ [15] ---------- (6)}$$

It uses Harmonic mean instead of Arithmetic mean, thus pushing the values to their extreme ends. This helps the model to be more dependable/comparable [16].

4. RESULT

In this paper, we have used logistic regression and random forest algorithm and calculated f1 score and accuracy respectively for each algorithm. We found that majorly the students ranging from 21-23 aged group were in high stress (1) and also the ratio between male is to female in stress was found to be equal.

The performance parameters we considered above (in section III.C) were calculated as follows:

Table 2. Results Calculated for the performance parameters

Parameters	**Logistic Regression**	**Random Forest**
Accuracy	95.23%	90.47%
TPR	92.31%	91.66%

TNR	88.88%	88.88%
FPR	0%	11.11%
Precision	100%	91.66%
f1-score	0.9411	0.8888

We can conclude that our data analysis and implementation is performing well, giving an accuracy of 95.23% and test f1 score as 0.9411 using logistic regression

	precision	recall	f1-score	support
0	0.92	1.00	0.96	12
1	1.00	0.89	0.94	9
accuracy			0.95	21
macro avg	0.96	0.94	0.95	21
weighted avg	0.96	0.95	0.95	21

Fig 8. Classification report for Logistic Regression

and accuracy = 90.47% and f1 score = 0.8888 using random forest.

	precision	recall	f1-score	support
0	0.92	0.92	0.92	12
1	0.89	0.89	0.89	9
accuracy			0.90	21
macro avg	0.90	0.90	0.90	21
weighted avg	0.90	0.90	0.90	21

Fig 9. Classification report for Random Forest

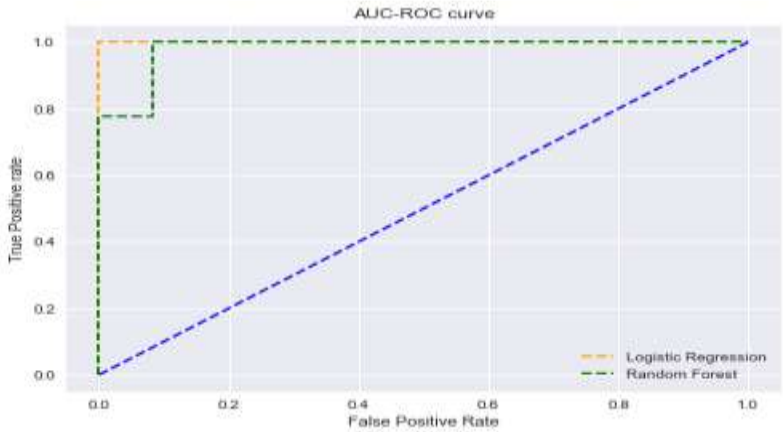


Fig. 10. Plot of AUC-ROC Curve

5. CONCLUSION

In this paper till now we have used two algorithms that are random forest and logistic regression [12] and after calculating we found that the results were good enough with accuracy of 95.23% and 90.47% respectively. Due to less availability of data, our model predicted almost everything correctly as 114 attributes were a key factor in model's performance. In future, our

main target will be to acquire more and more data so the precision of model becomes more and more sharp and also, we'll try to implement the data with various other models like Naive Bayes, SVM, and may-be we can try many other different validation techniques like k-fold, cross or may be even hybrid, which might improve results.

REFERENCES

[1]. C. Goumopoulos and E. Menti, "Stress Detection in Seniors Using Biosensors and Psychometric Tests," *Procedia Computer Science*, vol. 152, pp.18-27, 2019.

[2]. R. Ahuja and A. Banga, "Mental stress detection in university students using machine learning algorithms," *Procedia Computer Science*, vol. 152, pp. 349-353, 2019.

[3]. T. Kovilpatti and V. Kalaivani, "Analyzing social media data for understanding students learning experiences and predicting their psychological pressure," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 7, pp. 513-521, 2018.

[4]. Y. C. Tapidingan and D. Paseru, "Comparative Analysis of Classification Methods of KNN and Naïve Bayes to Determine Stress Level of Junior High School Students," *Indonesian Journal of Information Systems*, vol. 2, no. 2, pp. 80-89, 2020.

[5]. G. Verma and H. Verma, "Model for predicting academic stress among students of technical education in India," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 4, 2020.

[6]. P. S. Behere, R. Yadav, and P. B. Behere, "A comparative study of stress among students of medicine, engineering, and nursing," *Indian journal of psychological medicine*, vol. 33, no. 2, pp. 145-148, 2011.

[7]. M. Goran-Stanišić, M. Michalak, and A. Posadzy-Małaczyńska, "Drinking alcohol as a way of coping with stress in students of medical faculties," *Psychiatr. Pol*, vol. 54, no. 2, pp. 265-277, 2020.

[8]. K. A. Kalmakis et al., "Adverse childhood experiences, post-traumatic stress disorder symptoms, and self-reported stress among traditional and nontraditional college students," *Journal of American college health*, vol. 68, no. 4, pp. 411-418, 2020.

[9]. J. Tolles, and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes," *Jama*, vol. 316, no. 5, pp. 533-534, 2016.

[10]. D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[11]. N. Kardani et al., "Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data," *Journal of Rock Mechanics and Geotechnical Engineering*, vol.13, no.1, pp. 188-201, 2021.

[12]. S. Kour, R. Kumar and M. Gupta, "Analysis of student performance using Machine learning Algorithms," *Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021) DVD Part Number: CFP21N67-DVD*; ISBN: 978-0-7381-4626-3, 2-4, September 2021. (Indexed: Scopus) Available at: https://ieeexplore.ieee.org/document/9544935.

[13]. P. Sharma, R. Kumar, and M. Gupta, "Impacts of Customer Feedback for Online-Offline Shopping using Machine Learning," In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1696-1703). IEEE, Oct 2021.

[14]. M. Gupta, R. Jain, A. Gupta, and K. Jain, "**Real-time analysis of Covid-19 Pandemic on Most populated countries worldwide**," CMES-Computer Modeling in Engineering & Sciences, This article belongs to this Special Issue: Computer Modelling of Transmission,

Spread, Control and Diagnosis of COVID-19). ISSN: 1526-1506, 14th Sept 2020, DOI:10.32604/cmes.2020.012467.

[15]. T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp.861-874, 2006.

[16]. M. Gupta and D. Dahiya, "Performance Evaluation of Classification Algorithms on Different Data Sets", Indian Journal of Science and Technology, vol. 9, no. 40, pp. 1-6, DOI: 10.17485/ijst/2016/v9i40/99425, Oct. 2016. ISSN (Print): 0974-6846 ISSN (Online):0974-5645.

[17]. M. Gupta, V. Kumar-Solanki and V. Kumar-Singh, "A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity", IngenieríaSolidaria, vol. 13, no. 21, pp. 37-44, April, 2017.doi:10.16925/issn.1900-3102.