Machine Learning Model for Prediction of Stress Levels in Students of Technical Education

Garima Verma*, Sandhya Adhikari, Vaishnavi Khanduri, Shubhi Tandon, Shubhadika Rawat, Palak Singh

DIT University, Dehradun, INDIA

* Corresponding author's email: Garimaverma.research@gmail.com

doi: https://doi.org/10.21467/proceedings.100.5

ABSTRACT

An alarming rate of teenagers and youths are facing depression and anxiety now a days. One of the major reasons is the mental stress. The objective of this study is to identify the factors that affects the mental condition like depression, stress, anxiety in the students studying at the college level specially in engineering colleges. Two machine learning models logistic and Support Vector Machine (SVM) are proposed to predict the stress level of the students. For this study the dataset of 513 students has been collected by some engineering colleges of the northern India studying at graduation level. The data is collected using online and offline questionnaires. Accuracy, precision, recall and AUC-ROC curve performance metrics are being used to measure the performance of the models. The accuracy achieved by the logistic regression is 67% while the SVM has achieved 86.84%.

Keywords – Support vector machine, Logistic Regression, Stressed, Unstressed.

I. INTRODUCTION

Depression and anxiety have become a very common problem in youngsters now a days. Specially those who started their college life just after the completion of school. The changes in life style, college atmosphere, study pressure, career pressure, different pattern of teaching, increase work load, hostel life etc. can be various reasons for the development of mental stress. Generally above factors are the situations which bring prompt changes in the life of a student. Slowly the effect of these factors become critical in their mind and start creating stress [1]. Also, the increase stress level starts damaging their health and mental peace. There are various cases that comes in front of us in day to day life, where students commit suicide or starts taking drugs [2]. This study is an effort to analyze the stress level of the students at the early stage, so that different type of precautions can be taken by their parents, teachers, friends, college management etc.

The paper is further divided in 6 sections. Section 2 presents a literature survey of existing state of arts. Section 3 presents methods and material used for this work such as machine learning models, data set description etc. Section 4 presents the proposed model. Section 5 describes the performance of the model using performance metrics. Finally, the conclusion and future scope of the work is presented in Section 6.

II. RELATED WORK

Now a day's stress among students is a common scenario. Earlier terms like "stress" or "stressful lives" were meant to be for people busy with their jobs or people who had to think about maintaining financial stability, kids' future and similar worries. But nowadays it is quite common to see teenagers suffering from stress and other serious stress related problems like- anxiety, high blood pressure and even depression [1],

Proceedings DOI: 10.21467/proceedings.100; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-942709-6-6 (eBook)

[2]. There are no fixed reasons which can transform one’s life to a stressful life. During this research study we communicated with various university students of studying different courses at different levels specially to engineering students. We found some reasons which may lead to stress like- peer pressure, pressure of maintaining attendance, lack of cooperation, difficulties in making friends and changing to new environment etc. The applications of machine learning methods plays very vital role in the prediction of stress. There are different studies done by various researchers in this field such as Xu et al [3], Hussain et al [4], Ahuja et al [5], and Mohd. et al [6], for measuring and analyzing stress in human by different methods.

Xu et al [3], proposed a model based on cluster analysis to measure stress. The model used physiological signals extracted by the testing of various type of stress given to the different subjects (human) in different circumstances. Further they have used two techniques for analysis- k-means for dividing dataset into clusters and the neural network for stress evaluation. Hussain et al [4], proposed a model for prediction of student performance using machine learning algorithms and the need of additional assistance by teacher or mentor has been analyzed on the data of technology-enhanced learning (TEL) system called digital electronics education and design suite (DEEDS). They have applied Five-fold cross-validation and random division of the data into portions to assess the performance of the models. The machine learning algorithm used in the model are Artificial neural network (ANN), logistic regression (LR), Naive bayes classifiers (NBC), support vector machine (SVM), decision tree (DT) resulting accuracy of 75%, 73%, 75%, 75%, and 69% respectively. They have also applied feature selection method using Alpha-investing to the previous models and attained an accuracy of up to 80% with SVM model, which is highest among all. Ahuja et al [5], proposed machine learning models for calculating the mental stress during exam and while usage of internet. Correlation of the stress with the usage of internet is drawn with the dataset of 206 students’ data of Jaypee Institute of Information Technology. The accuracy and performance is improved with 10-fold cross validation applied on classification algorithms: Random Forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbour with the accuracy and found accuracy 83.33%, 71.42%, 85.71%, and 55.55 respectively. Mohd et al [6], proposed a model for prediction of students’ depression on data set of UniKL MIIT (September – December 2016 semester). Authors collected data by a questionnaire containing demographic information, student stress (interpersonal factor, intrapersonal factor, environment factor and academic factor), and CES-D: Center for Epidemiologic Studies Depression Scale). The calculation for students’ depression is taken from CES-D. In the study authors performed the performance comparison between Multivariate Logistic Regression (LR) and Multilayer Perceptron Neural Network (MLPNN) with Back propagation algorithm (BP), resulting in the accuracy of 62.5% and 71.8% respectively. The problem with all the above discussed work is that, the authors have considered all the factors which can generate stress in a student generally. But no work has considered the scenarios of mental stress in students, newly admitted in graduation for technical education. This work is focused on the students taken admission in technical education after completion of 12th from school, where the students have no exposure of technical teaching and learning environment.

III. METHODS AND MATERIALS

A. Support Vector Machine (SVM)

SVM is a machine learning approach, which is specially used for classification and regression problems. It is capable of handling continuous as well as categorical data. SVM represents, data points which are represented on space of hyperplane, can be categorized into two groups. Points which has similar properties comes under same group [7]. In SVM with linear kernel the dataset is represented as p-dimensional vector that can be separated by p-1 planes called as hyper-planes. These planes are used to set the boundaries between data groups. According to the distance between two classes, suitable hyper-plane is selected.

Let given a data points of training dataset, which are n, can be defined as in Eq. 1. –

$$(\overrightarrow{x_1}, y_1) \dots\dots\dots (\overrightarrow{x_n}, y_n) \qquad (1)$$

Where y1, represents a class of x1, and the value of y1 can be 1 or -1, here x1 is a real vector [7], [8].
Maximum-margin hyper-plane defined as Eq. 2 –

$$\overrightarrow{w}, \overrightarrow{x}, -b = 0 \qquad (2)$$

Where w is a normal vector and $\frac{b}{\|\overrightarrow{w},\|}$ is offset of hyper-plane along $\overrightarrow{w}$.

The SVM algorithm in practice is used in the form of a kernel. In this study linear kernel SVM has been used. Linear kernel is a dot product and defined in Eq. 3-

$$K\left(x_{,}\, x_i \right) = sum(x * x_i) \qquad (3)$$

B. Logistic Regression

Logistic Regression is a method used by Machine Learning from the field of insights. The method helps in solving classification problems. The attempt, to predict output variable Y from the given set of X inputs, is done by Logistic regression and linear regression essentially which are the same. These are form of supervised learning methods, which attempts to anticipate the reactions of unlabeled, unseen data by first training with labeled data, a lot of perceptions of both independent(X) and dependent (Y) variables. It is a procedure used to model and analyze the relationship between variables and often times how they contribute and are related to producing a specific result together [9]. The probability that Y, the response variable belongs to a certain category will be modeled by logistic regression. The response variable in many cases will be binary one, so logistic regression will want to model a function y=f(x) that outputs a standardized value that ranges from say 0 to 1 for all values of X corresponding to the two possible values of Y. This is done using the Logistic Function [9], [10].

C. Description of Dataset

For preparing the dataset, the data has been collected using questionnaire and interviews. For collection of data online and offline both modes has been used. The data captured by the questionnaire contains 22 features such as demographic information, atmospheric information such as –whether, health, surroundings etc. apart from these other information such as Lack of Support from Friends, Attendance Pressure, Debar Pressure, Lack of proper Information, language problem, etc. used as independent features and stress is taken as the target variable. Dataset contains total 513 cases and among all 274 cases recorded as unstressed and remaining were stressed cases shown in Fig. 1. The statistical description of some factors of the dataset is shown in Table -1.

TABLE I: DESCRIPTION OF DATASET

	HeavyWorkload	***LackofSupportFM***	***LackofSupportParents***	***LackofSupportFriend***	***AttendancePressure***	***DebarPressure***	***LackofInfo***	***Timings***
Mean	3.513513514	3.305555556	2.027777778	2.555555556	4.513514	4.324324	3.722222	4.166667
Median	4	4	1	2	5	5	4	4
Mode	4	4	1	1	7	2	4	4
Std. Deviation	1.660348384	1.670234277	1.521016786	1.697804371	2.103194	2.224286	1.750283	2.077086
Sample Variance	2.756756757	2.78968254	2.313492063	2.882539683	4.423423	4.947447	3.063492	4.314286
Kurtosis	-0.222719553	0.074194828	1.29127115	0.72789576	-1.21559	-1.57276	-1.0882	-1.19567
Skewness	0.263688711	0.496498922	1.549646037	1.160134405	-0.35141	-0.16159	-0.15728	-0.21493
Range	6	6	5	6	6	6	6	6
Minimum	1	1	1	1	1	1	1	1
Maximum	7	7	6	7	7	7	7	7
Count	37	36	36	36	37	37	36	36

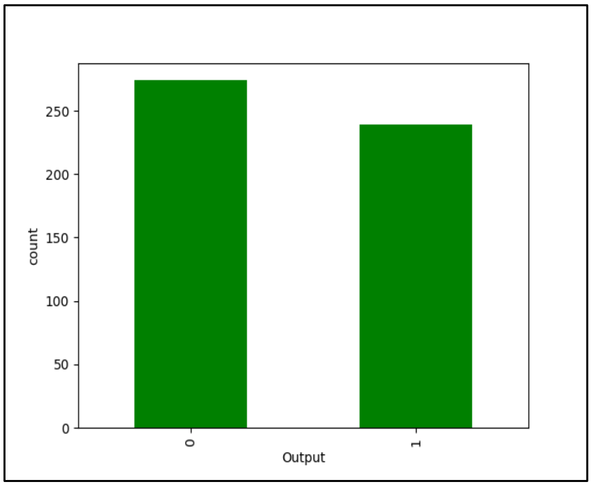

Fig. 1 Plot between unstressed and stressed students

IV. PROPOSED MODEL

Proposed model algorithm and flow chart is shown in Fig. 2 and 3 respectively. The figure shows the flow and steps of the model. After loading of the dataset, preprocessing is done to make dataset better. Specially to handle missing values and wrong values in the dataset. For prediction logistic regression and kernel based Linear SVM algorithm is used. Both algorithms are mostly used for classification problems. There are various kernel methods can be used with SVM, but in this study linear kernel is used for the prediction of the stress. The dataset was divided in training and test set. 80% of data is taken as training and remaining 20% is taken as test dataset. The model performance was evaluated using confusion matrix, Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve, precision, recall and accuracy.

```
Begin
    1. Load dataset
    2. Data cleaning and refining
    3. Define X as collection of independent variables
    4. Define Y as dependent variable or target
    5. Split dataset in training and testing data
    6. Creation of Linear SVM and Logistic Regression model
    7. Calculate performance metrics
End
```

Fig. 2 Algorithm for Proposed model

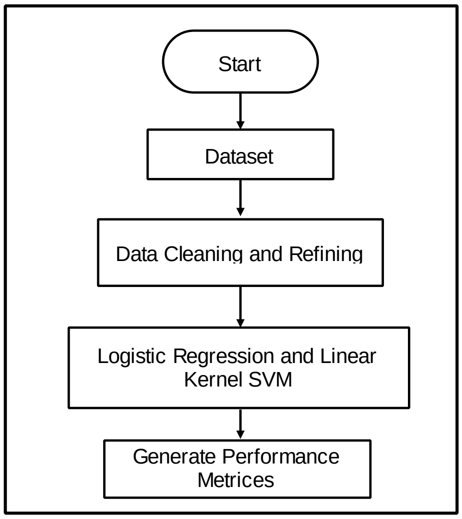
Start
Dataset
Data Cleaning and Refining
Logistic Regression and Linear Kernel SVM
Generate Performance Metrices

Fig. 3 Flow chart of Proposed model

V. RESULTS

A. Confusion Matrix

Confusion matrix is a table, which is very commonly used to describe the performance of the classification models. The table has 4 types of value- True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN). TP represents all the cases which are predicted as true and actually they are true [11]. FP represents all cases where model predicted true, but actually they are not. FN represents the cases where the model predicted no, but actually they have stress, TN represents cases where model predicted no, and actually they do not have stress also. Confusion matrix of the proposed model is shown in Table 2.

In this table there are two possible predicted classes: "1" and "0". We have predicted the stress level of students where, "1" means they are stressed, and "0" means they are not stressed. Out of these 513 students, the classifier predicted “1” 75 times, and “0” 79 times. In reality, 73 students are stressed, and 81 are not stressed.

TABLE II: CONFUSION MATRIX

		Predicted	
		Testing Data	
		0	1
Observed	0	52	29
	1	27	46

B. Precision, Recall and Accuracy

Precision is the percentage, when model is making prediction how frequently it is giving correct results. In the proposed model the precision is 61.33 % for logistic regression and 70.3% for SVM. For calculating the precision Eq. 4 has been used.

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \qquad (4)$$

Recall represents the percentage value, if there are students who are stressed in test dataset and proposed model can identify that. For the proposed model recall is 63.01% for logistic regression and 74.02% for SVM. Recall has been calculated by Eq. 5.

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

Over all accuracy of the model is predicted approximately 63.64% for logistic regression and 78% for SVM. Accuracy has been calculated by Eq. 6.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

C. AUC-ROC Curve

AUC-ROC curve is used to check the performance of the classification model. This metrics is used to visualize the performance of the algorithm in the form of a graph. ROC is basically a curve of probability and degree is measure by AUC curve. The model gets better and better according to the high values of AUC [12]. ROC curve is plotted between True positive rate (TPR) and False Positive Rate (FPR). ROC curve is shown with score of proposed model is Fig. 4 and 5. The TPR and FPR has been calculated according to the Eq. 7, and 8 respectively. A good model always has AUC near to 1.

$$TPR = \frac{TP}{TP+FN} \qquad (7)$$

$$FPR = \frac{FP}{TN+FP} \qquad (8)$$

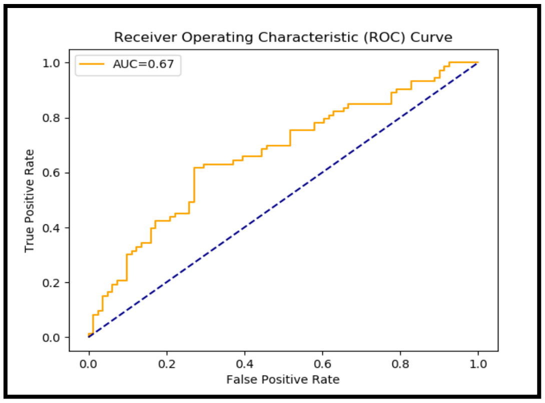

Fig. 4 ROC curve of proposed model using test dataset

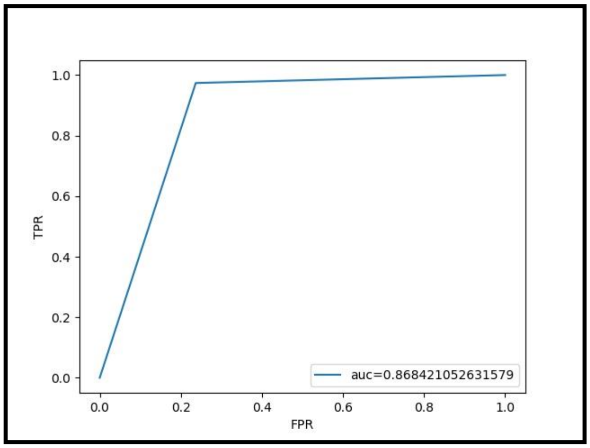

Fig. 5 ROC curve of proposed model using test dataset

VI. CONCLUSION AND FUTURE SCOPE

In this study an effort has been made to develop a model that could predict stress level in students starting technical education just after 12th in the various Universities. The model uses Logistic Regression algorithm and Support Vector Machine based on linear kernel. Experiment was performed using the dataset of 513 students collected from various students by using online and offline questionnaire. Results of the experiments shown in the form of confusion matrix, precision, recall, overall accuracy and AUC-ROC curve. The AUC-ROC accuracy of the SVM model is 86.84% and Logistic model is 67%. In the future, the study may be improved by including more machine learning algorithms and by increasing the size of dataset, also by introducing some new features with the help of some feature engineering technique.

REFERENCES

[1] Towbes, L.C. and Cohen, L.H., 1996. Chronic stress in the lives of college students: Scale development and prospective prediction of distress. Journal of youth and adolescence, 25(2), pp.199-217.

[2] Ghaderi, A., Frounchi, J. and Farnam, A., 2015, November. Machine learning-based signal processing using physiological signals for stress detection. In 2015 22nd Iranian Conference on Biomedical Engineering (ICBME) (pp. 93-98). IEEE.

[3] Xu, Q., Nwe, T.L. and Guan, C., 2014. Cluster-based analysis for personalized stress evaluation using physiological signals. IEEE journal of biomedical and health informatics, 19(1), pp.275-281.

[4] Hussain, M., Zhu, W., Zhang, W., Abidi, S.M.R. and Ali, S., 2019. Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review, 52(1), pp.381-407.

[5] Ahuja, R. and Banga, A., 2019. Mental Stress Detection in University Students using Machine Learning Algorithms. Procedia Computer Science, 152, pp.349-353.

[6] Mohd, N. and Yahya, Y., 2018, January. A Data Mining Approach for Prediction of Students' Depression Using Logistic Regression And Artificial Neural Network. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication (p. 52). ACM.

[7] Evgeniou, T. and Pontil, M., 1999, July. Support vector machines: Theory and applications. In Advanced Course on Artificial Intelligence (pp. 249-257). Springer, Berlin, Heidelberg.

[8] Verma, G. and Verma, H., 2019. Predicting Breast Cancer using Linear Kernel Support Vector Machine. Available at SSRN 3350254.

[9] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.

[10] Verma, H. and Verma, G., 2019. Prediction Model for Bollywood Movie Success: A Comparative Analysis of Performance of Supervised Machine Learning Algorithms. The Review of Socionetwork Strategies, pp.1-17.

[11] Ting, K. M.: Confusion Matrix. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining, Springer, Boston, MA (First online), (2017)

[12] Fawcett, T.:An introduction to ROC analysis. Pattern recognition letters, 27, 861-874, (2006).