Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms

Rumana Rois (rois@juniv.edu)
Jahangirnagar University https://orcid.org/0000-0002-0751-7104
Manik Ray
Jahangirnagar University Department of Statistics
Atikur Rahman
Jahangirnagar University Department of Statistics
Swapan. K. Roy
Bangladesh Breastfeeding Foundation

Research article

Keywords: Mental health, decision tree, random forest, support vector machine, feature selection, confusion matrix, ROC curves, k-fold cross-validation

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-468708/v1

Version of Record: A version of this preprint was published at Journal of Health, Population and Nutrition on November 27th, 2021. See the published version at https://doi.org/10.1186/s41043-021-00276-5.

Original research article

Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms

Rumana Rois [1*], Manik Ray [2], Atikur Rahman [3], and Swapan. K. Roy [4]

[1] Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh

[*]Corresponding author's Email: rois@juniv.edu

[2] Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh

Email: manikrayju@gmail.com

[3] Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh

Email: arahman@juniv.edu

[4] Bangladesh Breastfeeding Foundation (BBF), Institute of Public Health, Dhaka, Bangladesh

Email: skroy1950@gmail.com

Abstract

Background: Stress-related mental health problems are one of the most common causes of the burden in university students worldwide. Many studies have been conducted to predict the prevalence of stress among university students, however most of these analyses were predominantly performed using the basic logistic regression model. As an alternative, we used the advanced machine learning approaches for detecting significant risk factors and to predict the prevalence of stress among Bangladeshi university students.

Methods: This prevalence study surveyed 355 students from twenty-eight different Bangladeshi universities using questions concerning anthropometric measurements, academic, lifestyles, and health-related information, which referred to the perceived stress status of the respondents (yes or no). Boruta algorithm was used in determining the significant prognostic factors of the prevalence of stress. Prediction models were built using decision tree (DT), random forest (RF), support vector machine (SVM), and logistic regression (LR), and their performances were evaluated using parameters of confusion matrix, ROC curves, and k-fold cross-validation techniques.

Results: One-third of university students reported stress within the last 12 months. Students' pulse rate, systolic and diastolic blood pressures, sleep status, smoking status, and academic background were selected as the important features for predicting the prevalence of stress. Evaluated performance revealed that the highest performance observed from RF (accuracy=0.8972, precision=0.9241, sensitivity=0.9250, specificity=0.8148, AUC=0.8715, k-fold accuracy=0.8983) and the lowest from LR (accuracy=0.7476, precision=0.8354, sensitivity=0.8250, specificity=0.5185, AUC=0.7822, k-fold accuracy=07713) and SVM with polynomial kernel of degree 2 (accuracy=0.7570, precision=0.7975, sensitivity=0.8630, specificity=0.5294, AUC=0.7717, k-fold accuracy=0.7855). The RF model perfectly predicted stress including individual and interaction effects of predictors.

Conclusion: The machine learning framework can be detected the significant prognostic factors and predicted this psychological problem more accurately, thereby helping the policy-makers, stakeholders, and families to understand and prevent this serious crisis by improving policy-making strategies, mental health promotion, and establishing effective university counseling services.

Keywords: Mental health, decision tree, random forest, support vector machine, feature selection, confusion matrix, ROC curves, and k-fold cross-validation.

Introduction

Stress isn't a psychiatric diagnosis, but it's closely linked to mental health conditions including depression, anxiety, psychosis and post-traumatic stress disorder [1]. Stress can be defined as, "the inability to cope with a perceived (real or imaginary) threat to one's mental, physical, emotional, and spiritual well-being which results in a series of physiological responses and adaptations" [2]. This threat can be either positive (eustress) such as graduation or starting a new relationship, or negative, also called distress, with examples including academic probation or not being able to pay for school [3]. Students attending a university can experience both eustress and distress in the chronic (such as multiple roles and inadequate finances) or life event (such as relocation and death) forms [3]. The university years of an individual are emotionally and intellectually more demanding than almost any other stage of education [4]. At this stage, an individual faces a great deal of pressures and challenges that pose a variety of physical, social and emotional difficulties [5].

During this transitional period, students need to cope with the academic and social demands that they encounter in university studies that help in their preparation for professional careers by the acquisition of professional knowledge, transferable skills, and evidence-informed attitudes [6-9]. According to a national health college survey of National Mental Health Association (NMHA), 1 in 10 college students have been diagnosed with depression [10]. The

latest 2014 American College Health Association (ACHA) report [11] indicated that approximately half of students reported more than average or tremendous stress within the last 12 months. Moreover, scaling up mental health services will contribute to achievement of 1 of the targets of the Sustainable Development Goals (SDGs), endorsed at the United Nations General Assembly in 2015: by 2030, to reduce by one third premature mortality from noncommunicable diseases through prevention and treatment and promote mental health and well-being [12].

A plethora of research has focused on study of the prevalence of mental health problems among university population and the findings suggest that throughout the world, a substantial number of university students experience mental health problems [4, 6-9, 13-23]. In Bangladesh, there is much work in the literature regarding the prevalence of mental health problems among university students and the results emphasize that the prevalence of depression, anxiety, and stress has been reported to be as high as 54.3%, 64.8%, and 59.0%, respectively [9, 24-28].

Most stress-related studies have focused on the prediction of the prevalence of mental health problems using the logistic regression (LR) model. Prognostic modelling with LR is well-established, particularly for a dichotomous outcome. Although LR is a popular machine learning (ML) model for classification, we are interested to evaluate the performance of different ML models, including LR, to predict the prevalence of stress among Bangladeshi university students. ML in healthcare generally aims to predict some clinical outcomes on the basis of multiple predictors [29, 30]. The potential of ML in healthcare is vast, with demonstrations of ML-based tools being able to achieve human-level or above diagnostic and prognostic capabilities having been described in almost every clinical specialty [31]. The ML framework may explore more vital information on this crucial public health concern issue. Therefore, we are motivated to find the risk factors (features) and predict the prevalence of stress among Bangladeshi university students.

Materials and Methods

Participants and Procedures

We conducted a cross-sectional online-based study among undergraduate and graduate students of different universities of Bangladesh from January to March 2020, just before the COVID-19 outbreak in Bangladesh. The participants were included anonymously and voluntarily. Data were collected using convenience sampling via an online self-reported survey at the different universities throughout the country. Considering the 5% level of significance, 5% acceptable margin of error ($d = 0.05$), and $p = 0.363$ based on our pilot study (as 36.3% of university students reported stress within the last 12 months in our pilot study), the desired sample size has been estimated following the Cochran's formula:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{d^2}.$$

Hence, the required sample size was $n = 355.318 \approx 355$. Therefore, data from 355 participants were collected using a well-structured google form. Therefore, there were no incomplete questionnaires from any participants. The target variable, stress, was reported according to their perception of stress with a binary response (Yes=1, No=0). Input variables were included gender, academic year, their background (department) and university, and stress-related physical activity and lifestyle variables, such as sleep duration time, pulse rate (low= less than 60 beats per minute, normal= 60 to 100 beats per minute, high= more than 100 beats per minute), systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), drinking, and smoking habit. Students were classified according to WHO guidelines [32] as underweight (i.e., BMI < 20 kg/m^2), normal weight (i.e., 20 kg/m^2 < BMI < 25 kg/m^2), overweight/obese (i.e., BMI > 25 kg/m^2) based on their BMI value. For sleep duration, participants were asked to report the average duration of sleep per day as normal (6–7 hours), short (< 6 hours), or long (> 7 hours) [27]. According to the Joint National Committee report,

blood pressure (BP) categories were defined as Normotensive (normal BP) if the observed SBP was between 91 mmHg and 120 mmHg or DBP was between 61 mmHg and 80 mmHg; Prehypertensive if the observed SBP was between 121 mmHg and 139 mmHg or DBP was between 81 mmHg and 89 mmHg, and considered as Hypertensive if the observed SBP was equal to or above 140 mmHg and DBP was equal to or above 90 mmHg, and finally, Hypotension was defined as SBP being equal to or less than 90 mmHg or DBP being equal to or less than 60 mmHg [33-35].

Ethical issues

International ethical guidelines for biomedical research involving human subjects were followed throughout the study. After approval of the research proposal, ethical permission for data collection was received from the Department of Statistics, Jahangirnagar University, Bangladesh. The participants responded anonymously to the online survey by filling up an informed consent letter in the first section of the e-questionnaire. In the consent form, all the participants were provided with information concerning the research purpose, confidentiality of information, and right to revoke the participation without prior justification.

Statistical Analyses

This study aimed to classify and predict mental stress among Bangladeshi university students and assess the risk factors of their stress using different ML classification models, e.g., decision tree (DT), random forest (RF), support vector machine (SVM), and LR. Our methodology involves accordingly data collection and pre-processing, feature (the risk factors) selection using Boruta algorithm, splitting the entire data set into training and test data sets - applying ML models (DT, RF, SVM, LR) in the training data set and evaluate the performance of these models on the test data set, and finally using the best performed model predict mental stress based on the entire data set. The performances were evaluated using three performance parameters from the confusion matrix such as sensitivity, specificity, and accuracy, the area

under the receiver operating characteristics (ROC) curve (AUC), and the K-fold cross-validation. All ML models were performed using the scikit-learn module in Python programming language version 3.7.3. Only the Boruta algorithm was implemented to select the risk factors using the Boruta package in the R programming language [36].

Boruta Algorithm

Boruta algorithm was performed to extract the relevant risk factors for university students' perceived stress from this dataset. This is a wrapper build algorithm around the RF classifier to find out the relevance and important features with respect to the outcome variable [37].

Decision Tree (DT)

A DT is one of the most simple and intuitive techniques in ML, based on the divide and conquer paradigm [38]. A DT, whose internal nodes are tests (on input patterns) and whose leaf nodes are categories (of patterns), assigns a class number (or output) to an input pattern by filtering the pattern down through the tests in the tree [39]. Each test has mutually exclusive and exhaustive outcomes [39].

Random Forest (RF)

An RF algorithm has hyper-parameters specifying the number of trees and the maximum depth of each tree (effectively how many interactions are considered in the model), whereas the decision rules are the parameters [40]. The RF is an ensemble learning approach for classification using a large collection of decorrelated DT [41]. In this experiment, we have used 100 DT and Gini for impurity index to implement the RF algorithm in Python.

Support Vector Machine (SVM)

SVMs [42,43] are supervised learning methods that analyze data and recognize patterns. For a two-class learning task, an SVM training algorithm constructs a model or classification function that assigns new observations to one of the two classes on either side of a hyperplane,

making it a non-probabilistic binary linear classifier. An SVM model uses the kernel trick to map the data into a higher-dimensional space before solving the ML task as a convex optimization problem [41-44]. New observations are then predicted to belong to a class based on which side of the partition they fall. Support vectors are the data points nearest to the hyperplane that divides the classes [41]. We examined SVM models using the polynomial kernel of degree 2 and the linear kernel for this analysis.

Logistic Regression (LR)

LR is a probabilistic statistical classification model that predicts the probability of the occurrence of an event [41]. LR models the relationship between a categorical dependent variable and a dichotomous categorical outcome or feature. It is used as a binary (multiple) model to predict binary (multiple) responses, the outcome of a categorical dependent variable, based on one or more independent variables [38].

Confusion Matrix Performance Parameters

A confusion matrix provides a visual representation of actual versus predicted class accuracies [41]. To visualize the performance of the classification algorithm, it compares the predicted classification against the actual classification in the form of false positive, true positive, false negative and true negative information [41]. Therefore, the performance parameters are: accuracy is the number of data points correctly classified by the classifier, sensitivity is a measure of how well a classification algorithm classifies data points in the positive class, specificity is a measure of how well a classification algorithm classifies data points in the negative class, and precision is the number of data points correctly classified from the positive class [41].

Receiver Operating Characteristic (ROC) Curve

ROC curves offer another useful graphical representation for classifiers operating on datasets. Fawcett [45] provided a comprehensive introduction to ROC analysis, highlighting common misconceptions. The ROC curve shows the sensitivity of the classifier by plotting the rate of true positives to the rate of false positives. If the classifier is outstanding, the true positive rate will increase, and the area under the curve (AUC) will be close to 1 [38].

K-fold Cross-Validation

Cross-validation is a verification technique that evaluates the generalization ability of a model for an independent dataset [41]. It evaluates the performance of various prediction functions. In k-fold cross-validation, the training dataset is arbitrarily partitioned into k mutually exclusive subsamples (or folds) of equal sizes. The model is trained k times (or folds), where each iteration uses one of the k subsamples for testing (cross-validating), and the remaining k-1 subsamples are applied toward training the model. The k results of cross-validation are averaged to estimate the accuracy as a single estimation [41]. For this small sample size, we applied 3-fold, 5-fold, and 10-fold cross-validation techniques to evaluate the performance of classifiers.

Results

A total of 355 students have participated in this survey from 28 different universities throughout Bangladesh with the highest proportion of responses from Jahangirnagar University (56.1%), followed by the University of Dhaka (5.9%) and the University of Rajshahi (5.6%), detailed information is in the supplementary file. Among the participants, 204 were female (57.5%), 22.5% were overweight/obese, 15.8% were cigarette smokers, 8.5% were alcoholic, and 30.7% of university students reported stress within the last 12 months. The majority of the students had a normal pulse rate (76.9%), 63.4% were normal sleepers, 77.5% had normotensive BP for SBP and 76.9% had normotensive BP for DBP (Table 1). Just over half of the total sample, 62.3% (n=221) were masters students, followed by 13% (n=46) were first-

year university students. Highest proportion of participants 51.5% (n=183) were from science background, followed by 18.3% (n=65) were from arts. Stressed students were more likely to be male (35.1%), medical students (40%), first-year undergraduate students (41.3%), cigarette nonsmokers (39.3%), in low pulse rate (96.5%), normal sleepers (34.7%), overweight/obese (36.3%), had hypotension (100%) or hypertensive (100%) SBP, and had hypotension (100%) DBP as shown in Table 1.

Table 1 Frequency distribution and relationship with stress among university students

Variables	Total 355 n (%)	Stress (n = 109; 30.7%)		
		Yes (%)	χ^2	p-value
Gender				
Female	204 (57.5)	56 (27.5)	2.386	0.131
Male	151 (42.5)	53 (35.1)		
University				
1.Jahangirnagar University	199 (56.1)	59 (29.6)	38.811	0.066
2. University of Dhaka	21 (5.9)	5 (23.8)		
…	…	…		
27. National University	2 (0.6)	1 (50.0)		
28. University of Rajshahi	20 (5.6)	7 (35.0)		
Background				
Arts	65 (18.3)	20 (30.8)	2.891	0.576
Science	183 (51.5)	50 (27.3)		
Commerce	40 (11.3)	14 (35.0)		
Medical	30 (8.5)	12 (40.0)		
Engineering	37 (10.4)	13 (35.1)		
Academic Year				
1st Year	46 (13.0)	19 (41.3)	3.506	0.477
2nd Year	33 (9.3)	8 (24.2)		
3rd Year	31 (8.7)	10 (32.3)		
4th Year	24 (6.8)	8 (33.3)		
Masters	221 (62.3)	64 (29.0)		
Pulse Rate				
Low	57 (16.1)	55 (96.5)	200.75	0.000*
Normal	273 (76.9)	32 (11.7)		
High	25 (7.0)	22 (88.0)		
Alcoholic				
Yes	30 (8.5)	9 (30.0)	0.008	0.930
No	325 (91.5)	100 (30.8)		
Smoking Status				
Yes	56 (15.8)	22 (29.1)	2.301	0.129
No	299 (84.2)	22 (39.3)		
Sleep Time				
Less than normal	29 (8.2)	9 (31)	5.441	0.066
Normal	225 (63.4)	78 (34.7)		
More than normal	101 (28.5)	22 (21.8)		
SBP				
Hypotension	19 (5.4)	19 (100)	84.320	0.000*

Normotensive	275 (77.5)	59 (21.5)		
Prehypertensive	48 (13.5)	18 (37.5)		
Hypertensive	13 (3.7)	13 (100)		
DBP				
Hypotension	13 (3.7)	13 (100)	79.554	0.000*
Normotensive	273 (76.9)	63 (23.1)		
Prehypertensive	45 (12.7)	11 (24.4)		
Hypertensive	24 (6.8)	22 (91.7)		
BMI				
Underweight	77 (21.7)	24 (31.2)	1.710	0.425
Normal weight	198 (55.8)	56 (28.3)		
Overweight/obese	80 (22.5)	29 (36.3)		

*Statistically significant at the 0.05 level.

Table 1 also exhibits that stressed participants were significantly more likely than non-stressed participants to be in a low pulse rate (χ^2 =200.75, $p < 0.05$), had hypotension or hypertensive SBP (χ^2 = 84.320, $p < 0.05$), and had hypotension DBP (χ^2 =79.554, $p < 0.05$).

Features Selection

Figure 1 reveals that with the aid of the Boruta algorithm, six variables (Pulse rate, SBP, DBP, Sleep status, Smoking, Background [Dept]) were selected among ten surveyed variables as the risk factors to predict stress among Bangladeshi university students. Students' pulse rate, Sleep status, SBP, and DBP were the confirmed features and their smoking habit and background (Dept) were the tentative features for classifying their mental stress. Hereafter, these six variables were used to evaluate the performance of ML algorithms.

Machine Learning Models Evaluation

The performance of ML models such as DT, RF, SVM, and LR were evaluated using four performance parameters of the confusion matrix (Table 2), the area under the ROC curve (Figure 2), and the k-fold cross-validation approaches (Table 3). Considering 70% observations as the training data and 30% observation as the test data with the random seed 2379, using the scikit-learn module we estimated accuracy, sensitivity, specificity and precision of DT, RF, SVM, and LR algorithms to predict stress among university students and the results is illustrated in Table 2. Table 2 shows that the RF model was the efficient one to predict stress

among all the examined ML models based on the higher value of the performance parameters in all cases. For instance, the RF model provided 89.7% of accurate predictions (i.e., accuracy = 0.8972), 92.5% of positive cases that were predicted as positive (i.e., sensitivity = 0.9250), 81.5% of negative cases that were predicted as negative (i.e., specificity = 0.8148), and 92.4% of positive predictions that were correct (i.e., precision = 0.9241).

Table 2 Accuracy, sensitivity, specificity and precision of different ML models

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
DT	0.8785	0.9230	0.7586	0.9114
RF	**0.8972**	**0.9250**	**0.8148**	**0.9241**
SVM (Polynomial kernel)	0.7570	0.8630	0.5294	0.7975
SVM (linear kernel)	0.7570	0.8630	0.5294	0.7975
LR	0.7476	0.8250	0.5185	0.8354

Figure 2 illustrates the estimated AUC of DT, RF, SVM, and LR models, which were run using the scikit-learn module in Python 3.7.3 by considering 70% observations as training data and 30% observation as test data with the random seed 1439. To predict the prevalence of mental stress within the last 12 months among university students the estimated AUC was 0.8388, 0.8715, 0.7717, 0.6285, and 0.7822 using the ML models DT, RF, SVM with the polynomial kernel of degree 2, SVM with linear kernel, and LR, respectively. The RF algorithm performed better with the maximum AUC among all examined ML models. K-fold cross-validation was performed for 3-Fold, 5-Fold and 10-Fold repetitions with random seed 1 and shuffle argument 'True', and the results is organized in Table 3. The RF model performed better in 3-Fold, 5-Fold and 10-Fold cross validation based on the accuracy scores as shown in Table 3.

Table 3 Result of K-Fold cross validation of ML Models

Models	Accuracy (%) K-Fold		
	3-Fold	**5-Fold**	**10-Fold**
DT	0.8815	0.8901	0.8870
RF	**0.8844**	**0.8929**	**0.8983**
SVM (Polynomial kernel of Degree 2)	0.7718	0.7915	0.7855
SVM (linear kernel)	0.8085	0.8338	0.8309
LR	0.7830	0.7718	0.7713

To predict the mental stress within the last 12 months among Bangladeshi university students, the RF algorithm performed better based on the accuracy measure, the ROC, and the k-fold cross-validation approaches.

Model to predict stress

For the entire dataset, therefore, the best performed ML model, the RF model, was fitted to predict stress using the selected risk factors - Pulse rate, SBP, DBP, Sleep status, Smoking habit, and Background (department) of students, and the top one tree from the forest is visualized in Figure 3. All the nodes have five parts (feature's question, gini, samples, value and class) with a question based on a value of a feature, except the terminal leaf nodes have four parts (gini, samples, value and class) [46]. The part 'gini' indicates the Gini Impurity of the node, which is the average weighted Gini Impurity decreases as the path move down the tree, 'samples' is the number of observations in the node, 'value' is the number of samples in each class, and 'class' indicates the majority classification for points in the node ('class' is the prediction for all samples in the leaf node) [46].

Each feature's question has either a True (left nodes) or a False (right nodes) answer that splits the node. Based on the answer to the question, a data point moves down the tree and reaches a leaf node (the final decision). Moreover, the blue-type colored leaf indicates a prediction about stressed students and the orange-type colored leaf indicates a prediction about non-stressed students as shown in Figure 3. To predict any given student's data, simply move down the tree

in Figure 3, using the answer to the feature's question until arriving at a leaf node where the class is the prediction.

Table 4 organizes this decision path for five students' given data on their pulse rate, SBP, DBP, smoking habit, sleep status, and background (Dept) to predict their stress condition using the fitted RF model (Figure 3).

Table 4 Prediction of university student's stress using the fitted RF model

Pulse Rate	SBP	DBP	Smoking	Dept	Sleep Status	Predicted Stress
High	Hypertensive	Hypertensive	No	Arts	Normal	Stressed
Normal	Hypotension	Hypotension	No	Science	More than normal	Non-stressed
High	Normotensive	Hypotension	No	Medical	Normal	Non-stressed
Normal	Prehypertensive	Prehypertensive	Yes	Engineering	Less than normal	Stressed
Low	Normotensive	Hypotension	Yes	Medical	Less than normal	Stressed

LR analysis further revealed that cigarette smoker students were 4.112 times more likely (OR = 4.112, 95% CI = 1.591 – 10.628, $p < 0.05$) to be stressed than non-smokers. Respondents who had a normal pulse rate were less likely (OR = 0.002, 95% CI = 0.000 – 0.013, $p < 0.05$), and who had more than normal pulse rate were less likely (OR = 0.037, 95% CI = 0.004 – 0.389, $p < 0.05$) to be stress than those who had a low pulse rate (Table 5).

Table 5 Odds ratios (OR) with 95% CIs, and p-values obtained from the LR model

Variables	OR	(95% CI)	p-value
Pulse Rate			
Low (ref.)	1.000	-	-
Normal	0.002	(0.000 – 0.013)	0.000*
High	0.037	(0.004 – 0.389)	0.006*
Smoking			
No (ref.)	1.000	-	-
Yes	4.112	(1.591 – 10.628)	0.004*
Sleep Time			
Less than normal (ref.)	1.000	-	-
Normal	5.244	(0.811 – 33.911)	0.082
More than normal	5.660	(0.808 – 39.650)	0.081
SBP			
Hypotension (ref.)	1.000	-	-
Normotensive	0.000	(0.000 – .)	0.998
Prehypertensive	0.000	(0.000 – .)	0.998
Hypertensive	0.315	(0.000 – .)	1.000
DBP			
Hypotension (ref.)	1.000	-	-
Normotensive	0.000	(0.000 – .)	0.998
Prehypertensive	0.000	(0.000 – .)	0.999
Hypertensive	0.000	(0.000 – .)	0.999
Background			
Arts (ref.)	1.000	-	-
Science	1.428	(0.456 – 4.474)	0.541
Commerce	0.655	(0.117 – 3.682)	0.631
Medical	2.925	(0.545 – 15.702)	0.211
Engineering	3.210	(0.814 – 12.665)	0.096

Note: OR=1 for the reference category, * significant at 5% level

Discussion

University students are more vulnerable to stress and other mental health issues, which can negatively impact their health and academic performance [47-49]. The global prevalence of moderate to extremely severe levels is 60.8% for depression, 73% for anxiety, and 62.4% for stress [6-8, 18-22]. As a result, public concern for the mental health of university students has been rising and their stress has become a noticeable concept in public health. Motivated by such a noticeable public health concern, this research was conducted a prevalence study to find the risk factors and prediction of stress among university students in Bangladesh using different ML models. This prevalence study showed that one-third of university students reported stress within the last 12 months.

The study results reveal that university students' Pulse rate, SBP, DBP, Sleep status, Smoking status, and Background were the major risk factors for their stress using the ML features selection algorithm - Boruta. However, only students Pulse rate, SBP, and DBP were the significant factors for their stress using the conventional chi-squared test. Stressed students were more likely to be medical students (two-fifth), cigarette nonsmokers (3less than two-fifth), normal sleepers (more than one-third), in low pulse rate (less than one whole), had hypotension (exactly one-whole), or hypertensive (exactly one-whole) SBP, and had hypotension (exactly one-whole) DBP. Though stress and mental health differences exist between undergraduate and graduate students [50], the academic year was not a significant risk factor for our study. We observed that about two-fifths of the first year, followed by more than one-third of the fourth-year undergraduate students were stressed, whereas more than two-seventh graduate students were stressed. Gender was an insignificant risk factor for stress prediction, less than two-seventh of female students and more than one-third of male students were perceived stress within the last year. These findings of the current research have also differed from the earlier studies [4, 48, 51-53].

We evaluated the performance of ML models such as DT, RF, SVM, and LR to predict the stress of university students using four performance parameters of the confusion matrix, the AUC, and the k-fold cross-validation approaches. RF model was performed better to predict stress in all the situations with the highest 89.7% of accuracy, 92.4% of precision, 92.5% of sensitivity, 81.5% of specificity, 87.2% of AUC, more than 88% of accuracy in all the 3, 5, and 10-folds cross-validation techniques. The RF model was considered the individual and interaction effect of all the selected risk factors to predict the perceived stress of university students. Following the path in Figure 3, for any individual student with the given data, their perceived stress can be predicted as shown in Table 4. On the other hand, the LR model failed to estimate the confidence interval for the two significant predictor SBP and DBP, and

interpreted significantly only two predictor students' smoking status and pulse rate. Moreover, the RF model does not require any assumptions, whereas the popular classifier LR requires to fulfill all the underlying assumptions before estimating the model. Among all assumptions of the LR model, predictors independence and having a significant association with the outcome variable are the unavoidable assumptions. Therefore, this commonly used prognostic modeling is difficult to estimate properly. Furthermore, considering the high accuracy in prediction and better performance as well, other ML models such as the RF model will be informative and authentic (in terms of fulfilling the assumptions) to predict the perceived stress of university students.

Studies have also shown that mental health problems among university students are increasing in number as well as in severity [54]. Mental health problems can be a great source of psychological suffering and increase the risk of suicidal behaviors [6, 8, 18, 20, 55, 56]. Therefore, it is vital both to understand and then offer acceptable, effective, and accessible support for this potentially vulnerable group [57]. Counseling is the most consistently offered intervention and positive results have been demonstrated in services offering psychodynamic therapy, structured brief therapy, and integrative therapy [58, 59]. University counseling services in Australia, the UK, and the USA are reporting increases in help-seeking, with more students presenting with more severe problems [48, 60-62]. Although there is no noticeable awareness for university counseling services in Bangladesh, a reasonable number of researches carried out to address the prevalence of mental health problems among university students [9, 24-28, 63, 64], even during the COVID-19 pandemic [65-67].

The previous studies have been reported that the prevalence of stress among Bangladeshi university students as high as three-fifth of total respondents [9, 24-28]. However, our findings revealed that one-third of university students reported stress within the last 12 months. This lower prevalence rate of stress was observed as students were reported their last 12 months

feeling of stress by a binary response question (Yes or No), which is the foremost limitation of this study. Instead of using a binary response pattern, any structured scale such as depression, anxiety, and stress scale (DASS–21) with larger and more representative samples can be more informative to estimate the prevalence of stress. There is an increasing awareness of research to address the elevated risk of mental health problems in university students in Bangladesh, but a serious paucity of the health system, university counseling services, and policy of Bangladesh for supporting this potentially vulnerable group.

Conclusion

This study provides further evidence for the finding of elevated prevalence rate of stress among Bangladeshi university students. This psychological problem is very threatening as that can affect students' health, academic performance, and capacity to build their professional careers. Moreover, the magnitude of this problem needs to detect and understand, and hence, enable adequate and appropriate interventions for this vulnerable group. The ML framework can be detected the significant prognostic factors and predicted this psychological problem more accurately, thereby helping the policy-makers, stakeholders, and families to understand and prevent this serious crisis by improving policy-making strategies, mental health promotion, and establishing effective university counseling services.

Declarations

Ethics approval and consent to participate

Primary data were collected and participants were given no economic benefit, and anonymity was maintained to make sure the confidentiality and reliability of data. This study was conducted through online in full conformity with the international ethical guidelines for biomedical research on human participant research.

Consent for publication

All participants gave informed consent before taking part in the survey. They also provide their consent for publishing the analytical results from this survey without their identifiable information.

Availability of data and materials

The datasets that support the findings of this study are available on request.

Competing interest

The authors declare that they have no conflict of interest.

Funding

The authors received no specific funding for this work.

Author contributions

RR and AR jointly analysed data, drafted and reviewed the manuscript. RR conceived and supervised the study. MR collected data and performed the initial statistical analysis. SKR critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Authors are grateful to all participants of 28 universities students for giving their times voluntarily in the data collection during the survey.

References

1. Stress and our mental health: what is the impact & how can we tackle it? MQ Mental Health 2018. https://www.mqmentalhealth.org/stress-and-mental-health. Accessed May 16, 2018.
2. Seaward BL. Managing Stress: Principles and Strategies for Health and Well-Being (3rd ed.) 2002. Boston, MA: Jones and Bartlett Publishers.

3. Oswalt SB, Riddock CC. What to do about being overwhelmed: Graduate students, stress and university services. College Student Affairs Journal. 2007;27(1):24-44.
4. Saleem S, Mahmood Z. Mental health problems in university students: A prevalence study. FWU Journal of Social Sciences. 2013;7(2):124-130.
5. Rodgers LS, Tennison LR. A preliminary assessment of adjustment disorder among FirstYear College Students. Archives of Psychiatric Nursing. 2009;23(3):220-230.
6. Bayram N, Bilgel N. The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. Social Psychiatry and Psychiatric Epidemiology. 2008;43(8):667–672.
7. Kulsoom B, Afsar NA. Stress, anxiety, and depression among medical students in a multiethnic setting. Neuropsychiatric Disease and Treatment. 2015;11:1713–1722.
8. Haq UL, Dar MA, Aslam IS, Mahmood QK. Psychometric study of depression, anxiety and stress among university students. Journal of Public Health. 2018;26(2):211–217.
9. Mamun MA, Hossain MS, Griffiths MD. Mental health problems and associated predictors among Bangladeshi students. International Journal of Mental Health and Addiction. 2019;1-15.
10. National Mental Health Association. Finding hope & help: College student and depression pilot initiative, 2006. http://www.nmha.org/camh/college/index.cfm.
11. American College Health Association. American College Health Association-National College Health Assessment II: Reference group executive summary Spring 2014. https://www.acha.org. Accessed March 6, 2019.
12. World Health Organization. Investing in treatment for depression and anxiety leads to fourfold return, 2016. https://www.who.int/news/item/13-04-2016-investing-in-treatment-for-depression-and-anxiety-leads-to-fourfold-return.

13. Adewuya AO. Prevalence of major depressive disorder in Nigerian college students with alcoholrelated problems. General Hospital Psychiatry. 2006;28:169-173.
14. Nordin NM, Talib MA, Yaacob SN. Personality, Loneliness and Mental Health among Undergraduates at Malaysian Universities. European Journal of Scientific Research. 2009;36(2):285- 298.
15. Ovuga E, Boardman J, Wasserman D. Undergraduate student mental health at Makerere University, Uganda. World Psychiatry.2006;5(1):51-52.
16. Verger P, Guagliardo V, Gilbert F, Rouillon F et al. Psychiatric disorders in students in six French universities: 12-month prevalence, comorbidity, impairment and helpseeking. Social Psychiatry Psychiatric Epidemiology. 2009;45(2):189-199.
17. Seim RW, Spates CR. The prevalence and comorbidity of specific phobias in college students and their interest in receiving treatment. Journal of College Student Psychotherapy. 2010; 24:49-58.
18. Beiter R, Nash R, McCrady M, Rhoades D, Linscomb M et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. Journal of Affective Disorders. 2015;173:90–96.
19. Nadeem M, Ali A, Buzdar MA. The association between Muslim religiosity and young adult college students' depression, anxiety, and stress. Journal of Religion and Health. 2017;56(4):1170–1179.
20. Saeed H, Saleem Z, Ashraf M et al. Determinants of Anxiety and Depression Among University Students of Lahore. International Journal of Mental Health and Addiction. 2018;16(5):1283–1298.
21. Shamsuddin K, Fadzil F, Ismail WSW, Shah SA, Omar K et al. Correlates of depression, anxiety and stress among Malaysian university students. Asian Journal of Psychiatry. 2013;6(4):318–323.

22. Taneja N, Sachdeva S, Dwivedi N. Assessment of depression, anxiety, and stress among medical students enrolled in a medical college of New Delhi, India. Indian Journal of Social Psychiatry. 2018;34(2):157–162.
23. Bruffaerts R, Mortier P, Kiekens G, Auerbach RP et al. Mental health problems in college freshmen: Prevalence and academic functioning. Journal of Affective Disorders. 2018;225:97–103.
24. Alim SMA, Kibria HM, Islam SME et al. Translation of DASS 21 into Bangla and validation among medical students. Bangladesh Journal of Psychiatry. 2017;28(2):67–70.
25. Alim SMA, Rabbani HM, Karim MG, Mullick E et al. Assessment of depression, anxiety and stress among first year MBBS students of a public medical college, Bangladesh. Bangladesh Journal of Psychiatry. 2017;29(1):23–29.
26. Hossain MD, Ahmed HU, Chowdhury WA, Niessen LW, Alam DS. Mental disorders in Bangladesh: A systematic review. BMC Psychiatry. 2014;14(1):216.
27. Mamun MAA, Griffiths MD. The association between Facebook addiction and depression: A pilot survey study among Bangladeshi students. Psychiatry Research.2019;271:628–633.
28. Mamun MA, Rafi MA, Hasan M Z et al. Prevalence and psychiatric risk factors of excessive internet use among Northern Bangladeshi job-seeking graduate students: A pilot study. International Journal of Mental Health and Addiction. 2019. doi:https://doi.org/10.1007/s11469-019-00066-5.
29. Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ. Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence.2020;2(10):554-556.

30. Roberts M, Driggs D, Thorpe M, Gilbey J et al.Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 2021;3(3):199-217.
31. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine 2019;25(1):44-56.
32. World Health Organization. Addressing the socioeconomic determinants of healthy eating habits and physical activity levels among adolescents. 2006, Venice, Italy.
33. Chobanian AV, Bakris GL, Black HR. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: The JNC 7 report. JAMA. 2003;289(19):2560–2572.
34. Disease and conditions Index- Hypotension. National Heart Lung and Blood Institute. 2008. http://www.nhlbi.nih.gov/health/dci/Diseases/hyp/hyp_whatis.html Accessed on 2008 Sep 16.
35. Majed HT, Sadek AA. Pre-hypertension and hypertension in college students in Kuwait: a neglected issue. Journal of family and community medicine 2012;19(2):105.
36. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/; 2013.
37. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. Journal of Statistical Software. 2010;36(11):1-13.
38. Igual L, Seguí S.Introduction to Data Science. 2017, Springer, Cham.
39. Nilsson NL. Introduction to Machine Learning, 1997, CA.
40. Breiman L. Random Forests. Machine Learning. 2001;45(1): 5-32.
41. Awad M, Khanna R. Efficient Learning Machines, 2015. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_1

42. Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998;2(2):121-167.
43. Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. IEEE transactions on neural networks. 2001;12(2):181-201.
44. Vapnik VN. The Nature of Statistical Learning Theory 1995. Springer-Verlag, New York, Inc.
45. Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters. 2006;27:861–874.
46. Koehrsen W. An implementation and explanation of the random forest in Python. Medium, Towards Data Science, 31. 2018. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
47. Benton SA, Robertson JM, Tseng W, Newton FB, Benton SL. Changes in counseling center client problems across 13 years. Professional Psychology: Research and Practice. 2003;34:66-72.
48. Eisenberg D, Gollust SE, Golberstein E, Hefner JL. Prevalence and correlates of depression, anxiety, and suicidality among university students. American Journal of Orthopsychiatry. 2007;77:534–542.
49. Stanley N, Manthorpe J. Responding to students' mental health needs: Impermeable systems and diverse users. Journal of Mental Health. 2001;10(1):41- 52.
50. Wyatt T, Oswalt SB. Comparing mental health issues among undergraduate and graduate students. American journal of health education. 2013;44(2):96-107.
51. Mallinckrodt B, Leong FTL. Social support in academic programs and family environments: Sex differences and role conflicts for graduate students. Journal of Counseling and Development. 1992;70:716-724.

52. Sax LJ. Health trends among college freshmen. Journal of American College Health. 1997;45:252-262.
53. Hudd SS, Dumlao J, Phan E et al. Stress at college: Effects on health habits, health status and self-esteem. College Student Journal. 2000;34:217-238.
54. Hunt J, Eisenberg D. Mental health problems and help-seeking behavior among college students. Journal of Adolescent Health. 2010; 46:3-10.
55. Arafat SY, Mamun MAA. Repeated suicides in the University of Dhaka (November 2018): Strategies to identify risky individuals. Asian Journal of Psychiatry. 2019;39:84–85.
56. Shah M, Ali M, Ahmed S, Arafat SM. Demography and risk factors of suicide in Bangladesh: A six-month paper content analysis. Psychiatry Journal.2017;e5.
57. Brown JS. Student mental health: some answers and more questions, Journal of Mental Health. 2018;27(3):193-196.
58. Connell J, Barkham M, Mellor CJ. The effectiveness of UK student counselling services: An analysis using the CORE System. British Journal of Guidance and Counselling. 2008;36:1-18.
59. McKenzie K, Murray KR, Murray AL, Richelieu M. The effectiveness of university counselling for students with academic issues. Counsel Psychother Res, 2015;15:284–288.
60. Monk EM. Student mental health. Part 2: The main study and reflection of significant issues Counselling Psychology Quarterly. 2004; 17:33–43.
61. Avotney A. Students under pressure. American Psychological association, 2014;45(8):36.
62. Flatt AK. A suffering generation: Six factors contributing to the mental health crisis in North American higher education. College Quarterly. 2013; 16.
63. Hoque R. Major mental health problems of undergraduate students in a private university of Dhaka, Bangladesh. European Psychiatry. 2015;30(S1):1-1.

64. Arusha AR, Biswas RK. Prevalence of stress, anxiety and depression due to examination in Bangladeshi youths: A pilot study. Children and Youth Services Review. 2020;116.
65. Islam MS, Sujan MSH, Tasnim R, Sikder MT, Potenza MN, Van OsJ. Psychological responses during the COVID-19 outbreak among university students in Bangladesh. PloS one. 2020;15(12).
66. Faisal RA, Jobe MC, Ahmed O, Sharker T. Mental Health Status, Anxiety, and Depression Levels of Bangladeshi University Students During the COVID-19 Pandemic. International Journal of Mental Health and Addiction. 2021;1-16.
67. Ahammed B, Khan B, Jahan N, Shohel TA, Hossain T, Islam N. Determinants of Generalized Anxiety, Depression, and Subjective Sleep Quality among University Students during COVID-19 Pandemic in Bangladesh. Dr. Sulaiman Al Habib Medical Journal. 2021;3(1):27-35.

68. Uni Name	Uni Code	Frequency	Percent
JU =1	1	199	56.1
DU =2	2	21	5.9
HSTU =3	3	31	8.7
Shaheed Suhrawardy Medical College =4	4	10	2.8
Mymensingh Medical College =5	5	4	1.1
Noakhali Science and Technology University =6	6	2	.6
RUET =7	7	3	.8
AUST=8	8	3	.8
East West University=9	9	2	.6
Sher E Bangla Agricultural University=10	10	2	.6
Eden Mohila College =11	11	2	.6
Bangladesh University of professionals =12	12	3	.8
North South University =13	13	3	.8
University of Chittagong =14	14	4	1.1
Home Economics College=15	15	3	.8
Begum Rokeya University=16	16	14	3.9
Uiu=17	17	3	.8
DIU=18	18	2	.6
State University=19	19	6	1.7
Northen University=20	20	2	.6
Islamic University=21	21	3	.8
Tejgaon College=22	22	3	.8
BRAC University =23	23	3	.8

JnU=24	24	3	.8
NIP=25	25	1	.3
Rangpur Medical College =26	26	1	.3
National University =27	27	2	.6
RU=28	28	20	5.6

Table S1 Frequency of participated students from twenty-eight universities in Bangladesh

Figures

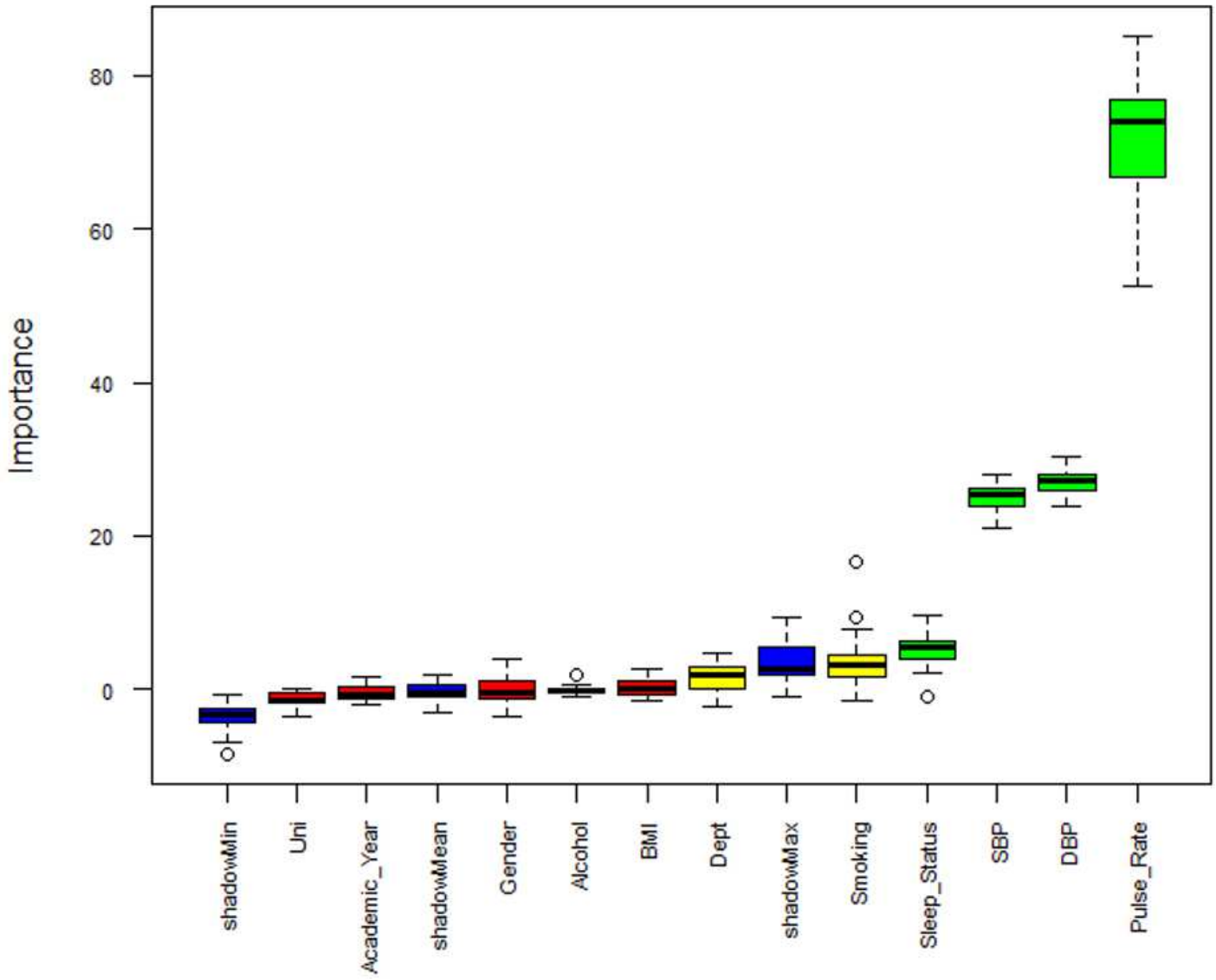

Figure 1

Features selection using the Boruta algorithm

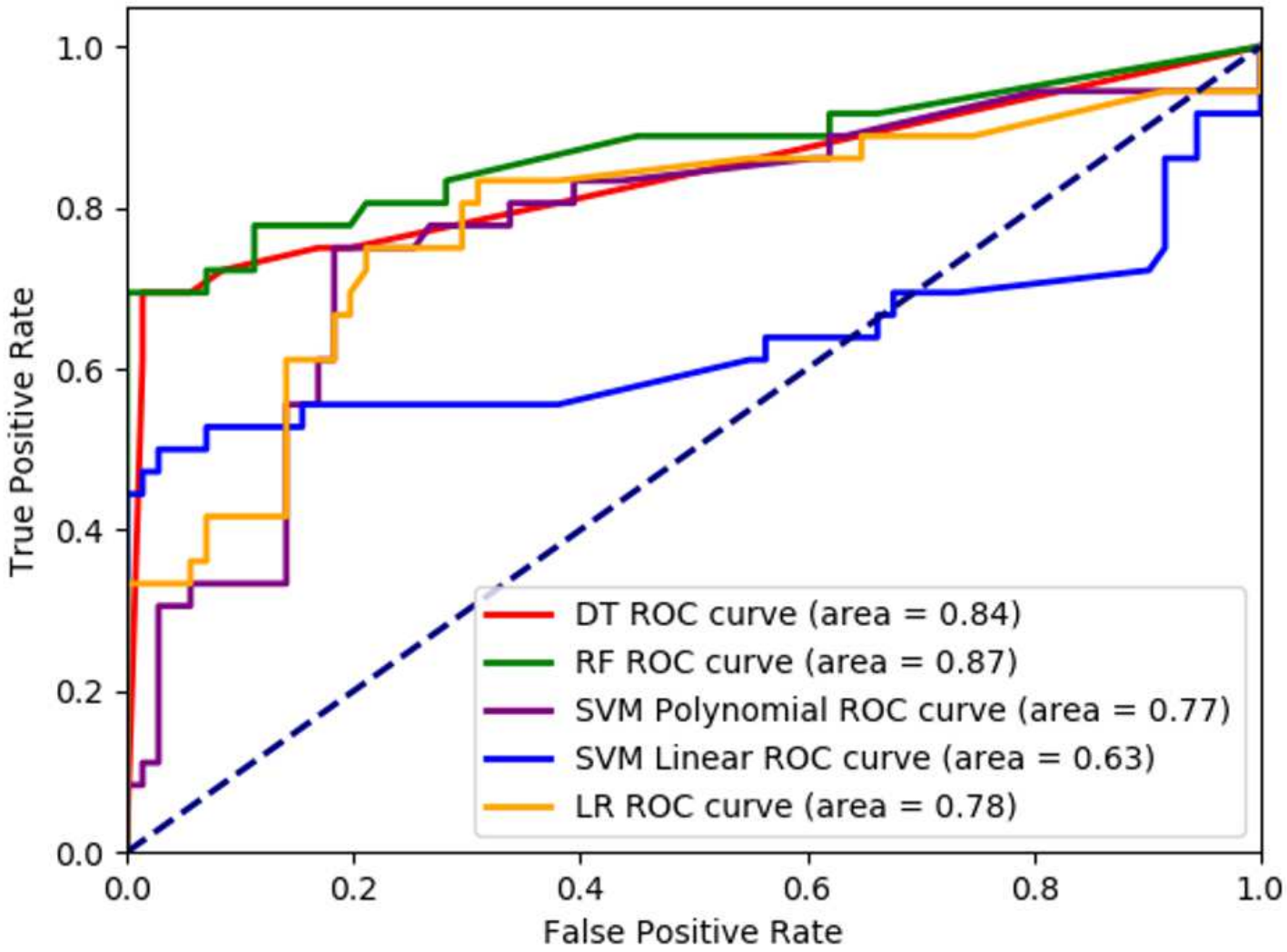

Figure 2

The ROC curves to predict mental stress using DT, RF, SVM, and LR models

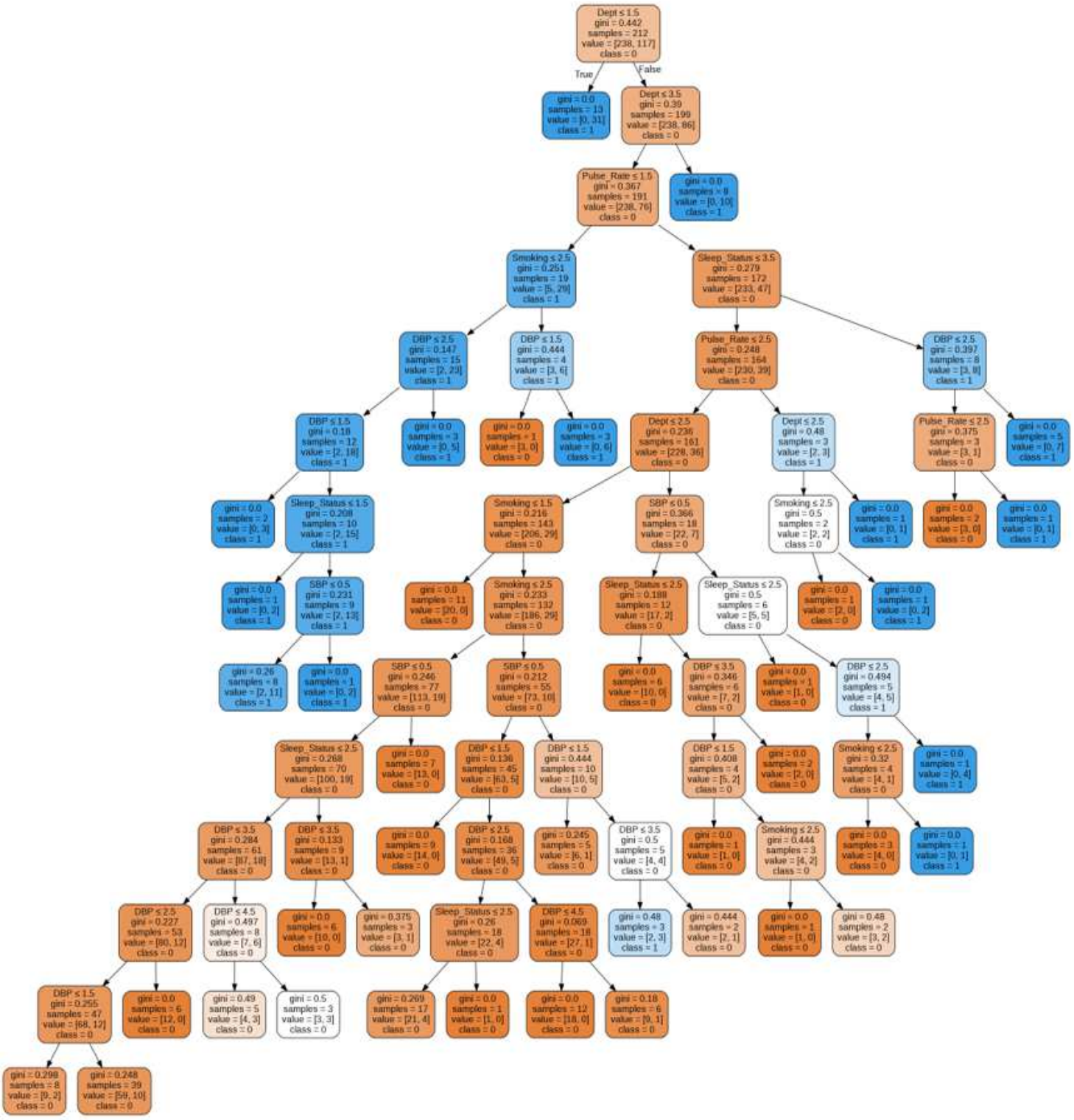

Figure 3

Top one tree from the fitted RF model to predict university student's mental stress