A Systematic Study of Student Stress Prediction in Education Sector Using Machine Learning

[1]Manoj Kumar V, [2]Mayur V, [3]Bharath M, [4]Khajabandur N M
Information and Science and Engineering, Malnad College of Engineering
Hassan-573202, India
Email id:manojkumarv724@gmail.com, abj@mcehassan.ac.in, mayurkainij@gmail.com, mbharath0326@gmail.com
mittalakodkhajabandu@gmail.com

***Abstract*—The mental well-being of students is increasingly important in educational institutions around the world. If the academic stress is not managed properly it will lead to various psychological problems affecting the overall performance of students. This paper surveys the use of ML techniques to predict and analyze student stress[1][2]. Various studies using ML techniques such as SVM, regression, random forests, decision trees, and artificial neural networks (ANNs) were reviewed to investigate their effectiveness in predicting students' stress levels based on different datasets and factors[11]. This review highlights the strengths and limitations of existing studies, identifies gaps in the literature, and makes gateway for further research to improve real-time stress prediction systems for students.**

***Keywords*—** *Student Stress, SVM, Regression, Decision Tree, Machine Learning, Education, Data Science, Academic, Results.*

I. INTRODUCTION

Students suffer from many mental illness like depression, pressure, stress, interpersonal sensitivity, fear, nervousness etc. Although many industries and companies provide programs related to mental health and try to alleviate the atmosphere in the workplace, the problem is far from under control. The prediction of stress in university students is one of the main and challenging tasks of the current education sector. Stress is considered the main thing used to create imbalance in each character's life, and is also considered the main issue for psychological adaptation and trauma reduction[3]. A number of studies deal with stress management in school students. Students who continue on to high school and college are suffering with their stress levels. Many times it can be decided to pay attention to lecturers as daily stimuli for a problem-free mind[9][6]. In order to reduce individual stress levels, many associations have been able to strengthen the complete stage of progress in monitoring the stage of stress in students and make sure they perform well in academies. Lack of stress administration will lead to some drastic injury that can sometimes completely affect education and can even cause extreme damage to students' fitness at various stages[2]. The individual family background was conceptualized as the main game that guided the path of our childhood. Children who live in the countryside or in cities are constantly looking at exclusive environments. Students with low placement usually have a low grade because of financial problems and family problems[14][15]. The performance of the faculty and students is structured mainly on both teaching faculties and home teaching. The current system is a manual process where it is difficult to identify stress in college students. There is no automation to predict student stress[11].

II. RELATED WORKS

This section summarizes and compares existing studies on student stress prediction using ML models. The studies reviewed focus on the algorithms, datasets, and methodologies used to classify and predict student stress.

A. Predicting Student Stress Levels Using Supervised Machine Learning and Artificial Neural Networks (ANN)

- **Authors:** S. Arya, Anju, and N. A. Ramli
- **Year:** 2024

Description:
People in various profession experience stress as a significant challenge, including students. This study identifies the primary stress factors affecting students at Tribhuvan University, Dharan, Nepal. By analyzing these factors, it aims to predict and prevent stress at early stages using advanced ML and deep learning techniques.
Methodology:
The study explores multiple guided learning techniques, including Vector Machine (SVM), Random Forest, Gradient Boosting, AdaBoost, CatBoost, LightGBM, ExtraTree, XGBoost, Logistic Regression, Decision Tree, Multilayer Perceptron (MLP), and Artificial Neural Networks (ANN).
Limitations:
The research is limited to a single university and considers a restricted number of stress-related factors, which may affect the generalizability of the findings.
Key Insights:
The SVM and ANN demonstrated high accuracy and effectiveness in predicting stress levels based on student demographics and lifestyle patterns.
Citation:
[1] S. Arya, Anju, and N. A. Ramli, "Predicting Student Stress Levels Using Supervised Machine Learning and Artificial Neural Networks," 2024.

B. Stress Detection in College Students Using a Machine Learning Algorithm

- **Authors:** Ms. Ancy Paul, Ms. Resija PR
- **Year:** 2024

Description:
Mental stress is a big concern, especially among youths. This study examines the effects of stress particularly on university

students at different stages of their lives, particularly exploring how factors such as social interactions, academic pressures, and screen time influence stress levels. The authors focus on the correlation between time spent on the internet and stress, aiming to identify different levels of stress among students. A dataset of 954 students from Vimala College (Autonomous), Thrissur, utilized, which was acquired via an online questionnaire where respondents were asked questions about their feelings in various situations. The dataset was analyzed using automated learning systems to classify stress levels into three categories: a) chronic, b) episodic, and c) acute.

Methodology:

The study employs a variety of guided learning techniques to predict stress levels, including Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), AdaBoost, Decision Tree, Stochastic Gradient Descent (SGD), and Linear Regression. These models were developed on the survey data, with the goal of categorizing the students into the three predefined stress classes.

Limitations:

The boundaries to this study are small dataset of 954 participants and restricts stress levels to just three parts, not capturing the full spectrum of stress experiences. Additionally, the study does not consider a wide range of psychological or surrounding conditions that could contribute to stress.

Key Insights:

The study concludes that ML techniques could successfully predict different levels of stress in students, especially using classifiers like Random Forest and SVM. However, the results were constrained by the limited number of stress-related factors considered in the analysis. It was observed that time spent on the internet had a significant correlation among students who had high stress.

Citation:

[2] Ms. Ancy Paul and Ms. Resija PR, "Stress Detection in College Students Using a Machine Learning Algorithm," 2024.

C. Predictive Analysis of Students' Stress Levels Using Machine Learning

- **Authors:** Dr. Anbarasi M, Sethu Thakkilapati, Veeragandham Rajeev Vas, Sarabu Venkata Bharath Viswath
- **Year:** 2023

Description:

Psychological pressure is a significant factor impacting the mental health of students, often leading to stress. The study identifies several causes of stress in student life, including academic pressures, social expectations, financial issues, and personal problems. Academic pressure, such as challenging grades, tight deadlines, and high expectations from teachers and peers, is highlighted as a major contributor to stress. The authors propose a method for educational institutions to predict student stress using ML techniques. The study focuses on using Random Forest (RF) and Decision Tree (DT) algorithms to predict stress levels among students.

Methodology:

The study utilizes two machine learning algorithms—Random Forest (RF) and Decision Tree (DT)—to predict stress levels in students based on various stress-related factors. The models are developed based on a dataset with features like academic performance, deadlines, social factors, and personal circumstances.

Limitations:

The study predicts stress without quantifying stress levels or considering a wider range of factors that may contribute to stress, such as mental health history or environmental factors. Additionally, the small size of the dataset may limit the precision and applicability of the models.

Key Insights:

The study analyzed that Random Forest and Decision Tree algorithms are best for predicting student stress. However, the effectiveness of these models was limited by the small dataset and the narrow scope of stress-related factors considered.

Citation:

[3] Dr. Anbarasi M, Sethu Thakkilapati, Veeragandham Rajeev Vas, and Sarabu Venkata Bharath Viswath, "Predictive Analysis of Students' Stress Levels Using Machine Learning" , 2023.

D. Predicting Stress in Indian School Students Using Machine Learning

- **Authors:** Aanchal Bisht, Shreya Vashisth, Muskan Gupta, Ena Jain
- **Year:** 2022

Description:

Children worldwide are experiencing significant stress, and if left unaddressed, this can have severe consequences. Stress can manifest in various ways, including anxiety, frustration, anger, or restlessness. This work aims to analyze the stress levels in Indian school students using ML techniques. The authors conducted a survey of more than 190 school children aged 14 to 18, asking 26 different questions to understand their levels of stress. The study employs multiple ML algorithms, including Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. The KNN model yielded the top score of 88% in predicting student stress levels.

Methodology:

The study uses a variety of ML models, including Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest, to determine the stress levels of school children. The dataset consists of responses to 26 questions regarding factors contributing to stress.

Limitations:

The study focuses solely on school children and does not account for a broader range of stress-related factors, such as socio-economic background, mental health history, or complex environmental variables, which may limit the generalizability of the results.

Key Insights:

The study demonstrated that KNN achieved the highest prediction accuracy (88%), highlighting the effectiveness of this model for predicting student stress. However, the study's scope was limited due to the narrow range of factors considered in the model.

Citation:

[4] Aanchal Bisht, Shreya Vashisth, Muskan Gupta, and Ena Jain, "Predicting Stress in Indian School Students Using Machine Learning,", 2022.

TABLE I. COMPARATIVE TABLE

Constraints	[1]	[2]	[3]	[4]
Less Parameters used	Yes	Yes	NO	Yes
Less Datasets used	No	Yes	Yes	Yes
Real Time Implementations	No	No	No	No
Dynamic Data	No	No	No	No
StressLevel Prediction	No	No	No	No
Solution Recc	No	No	No	No

III. METHODOLOGY

In the papers we referred the stress levels are predicted by collecting the stress related data from various sources. This data contains the parameters such as gender, age, financial problems, problems within the family, health issues, partial repair, pressure, regularity and interaction. In the next step, data preparation, the collected data is analyzed and only the relevant parts are extracted and segmented based on the requirements. This ensures that unnecessary data that could increase processing time is excluded. Subsequently, limitations are specified by identifying the parameters that will be used to analyze the stress level. It then uses supervised learning, a ML approach that depends on labeled training data with expected outputs. The system predicts stress levels using these parameters and ML algorithms. The model is evaluated with respect to accuracy by dividing the collected data into 90% training data and 10% test data. Finally, the results are visually represented, making it easier to interpret the outputs.

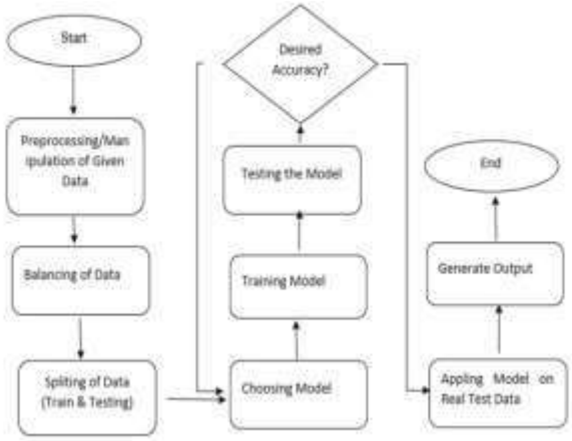

Start
Preprocessing/Manipulation of Given Data
Balancing of Data
Spliting of Data (Train & Testing)
Choosing Model
Training Model
Testing the Model
Desired Accuracy?
Appling Model on Real Test Data
Generate Output
End

Fig. 1. Methodology

IV. ALGORITHMS USED

4.1.1 SVM Algorithm:

Support Vector Machine (SVM) is a robust learning technique which is suitable to both linear and nonlinear classification tasks, as well as regression and anomaly detection[7]. Its flexibility allows it to be used in a many areas, including text categorization, image recognition, spam filtering, handwriting recognition, gene expression analysis, face detection, and the identification of outliers[3]. Due to its versatility, SVM is highly favored in many fields such as computer vision, bioinformatics, and artificial intelligence.

The equation for the linear hyperplane can be written as:

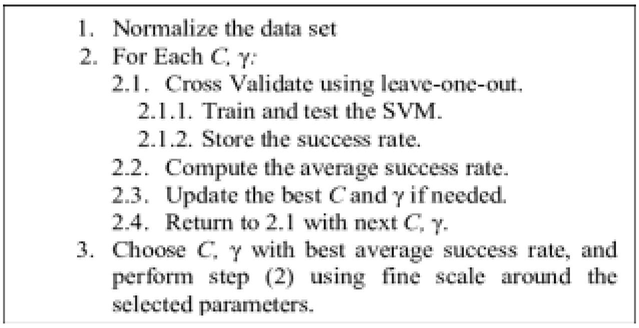
$$w^T x + b = 0$$

The vector W represents the normal vector to the hyperplane. i.e the direction perpendicular to the hyperplane. The parameter b in the equation represents the offset or distance of the hyperplane from the origin along the normal vector w.

The distance between a data point x_i and the decision boundary can be calculated as:

$$d_i = \frac{w^T x_i + b}{||w||}$$

where ||w|| represents the Euclidean norm of the weight vector w.



```
1. Normalize the data set
2. For Each $C$, $\gamma$:
   2.1. Cross Validate using leave-one-out.
      2.1.1. Train and test the SVM.
      2.1.2. Store the success rate.
   2.2. Compute the average success rate.
   2.3. Update the best $C$ and $\gamma$ if needed.
   2.4. Return to 2.1 with next $C$, $\gamma$.
3. Choose $C$, $\gamma$ with best average success rate, and
   perform step (2) using fine scale around the
   selected parameters.
```

Fig. 2. SVM Algorithm

4.1.2 Regression Algorithm

This section will focus on the development of pseudocode for implementing the linear regression method. By presenting the pseudocode, the goal is to simplify the implementation of the linear regression algorithm using high-level programming languages, making it more accessible for practical use[11].

Mathematically, we can represent a linear regression as:

Simple Linear Regression:

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y = \beta_0 + \beta_1 X$

where:

- Y is the dependent variable
- X is the independent variable
- β_0 is the intercept
- β_1 is the slope"

Multiple Linear Regression:

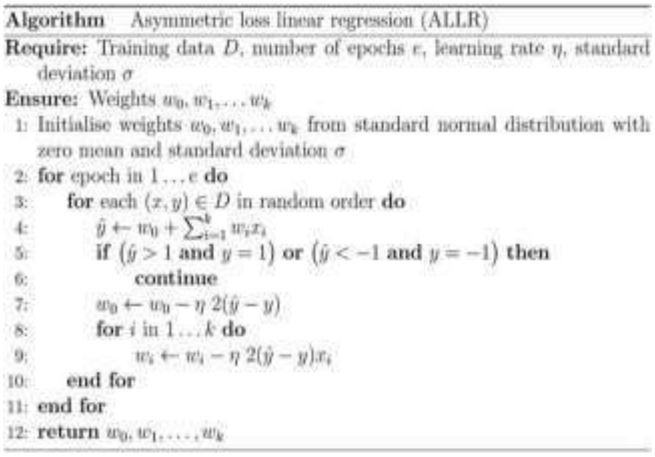
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

where:

- Y is the dependent variable

- $X_1, X_2, \ldots, X_n$ are the independent variables
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ are the slopes



```
Algorithm   Asymmetric loss linear regression (ALLR)
Require: Training data $D$, number of epochs $e$, learning rate $\eta$, standard
    deviation $\sigma$
Ensure: Weights $w_0, w_1, \ldots w_k$
Initialise weights $w_0, w_1, \ldots w_k$ from standard normal distribution with
    zero mean and standard deviation $\sigma$
for epoch in $1 \ldots e$ do
    for each $(x, y) \in D$ in random order do
        $\hat{y} \leftarrow w_0 + \sum_{i=1}^{k} w_i x_i$
        if $(\hat{y} > 1$ and $y = 1)$ or $(\hat{y} < -1$ and $y = -1)$ then
            continue
        $w_0 \leftarrow w_0 - \eta\, 2(\hat{y} - y)$
        for $i$ in $1 \ldots k$ do
            $w_i \leftarrow w_i - \eta\, 2(\hat{y} - y) x_i$
    end for
end for
return $w_0, w_1, \ldots, w_k$
```

Fig. 3. Regression Algoithm

4.1.3 Decision Tree Algorithm

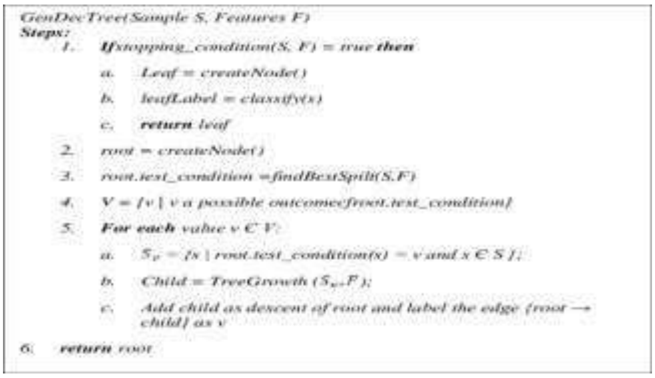
Decision trees are non-linear models that are highly flexible and powerful for both classification and predictive tasks. Unlike regression models, which uses a linear relationship between the input features and the target variable, decision trees do not have this restriction. As non-parametric models, decision trees are capable of capturing complex, non-linear relationships between predictors and targets[12], making them suitable for many applications where linear assumptions may not hold.

The mathematical formula for Decision tree is:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)
Where,
S= Total number of samples
P(yes)= probability of yes
P(no)= probability of no



```
GenDecTree(Sample S, Features F)
Steps:
    1.  If stopping_condition(S, F) = true then
        a.  Leaf = createNode()
        b.  leafLabel = classify(s)
        c.  return leaf
    2.  root = createNode()
    3.  root.test_condition = findBestSplit(S,F)
    4.  V = {v | v a possible outcomecfroot.test_condition}
    5.  For each value v ∈ V:
        a.  $S_v$ = {s | root.test_condition(s) = v and s ∈ S };
        b.  Child = TreeGrowth ($S_v$,F );
        c.  Add child as descent of root and label the edge (root →
            child) as v
6:  return root
```

Fig. 4. Decision Tree Algorithm

V. RESULTS

The difference in prediction performance of **Support Vector Machine (SVM)** and **Linear Discriminant Analysis (LDA)** models, showing the correct and incorrect predictions using a color-coded scheme. The **blue dots** represent correct predictions, while the **red dots** indicate incorrect ones. Both models display a combination of correct and incorrect predictions, but the SVM model appears to perform slightly better, as it shows a higher density of blue dots compared to LDA. In contrast, the LDA model seems to have a greater number of red dots, indicating a higher error rate. This visual comparison highlights the difference in prediction accuracy between the two models, suggesting that SVM might be more effective for the given dataset.

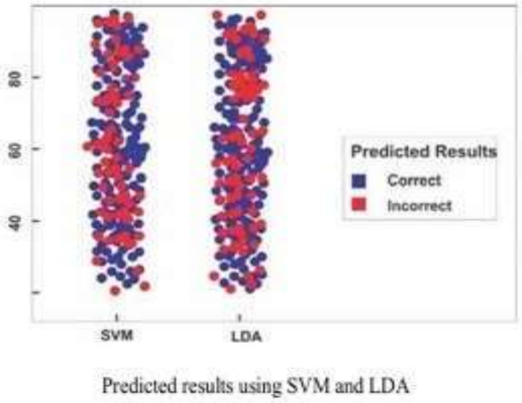


Fig. 5. Predicted Results

The bar chart compares the behaviour of various ML models based on their accuracy. The ANN achieved an accuracy of 74.69%, followed closely by Logistic Regression at 78.16% and SVM at 75.71%. The Gaussian Naive Bayes model performed the lowest with an accuracy of 50.59%, indicating poorer predictive performance. The Decision Tree and Random Forest models, though highly accurate at 96.2% and 98.4% respectively, are labelled as overfitted, showing that their behaviour may not generalize well to unseen data. This comparison highlights the trade-off between accuracy and generalizability, with Random Forest to be most accurate but at the risk of overfitting, while other models like Regression and SVM maintain a balance between accuracy and generalizability.

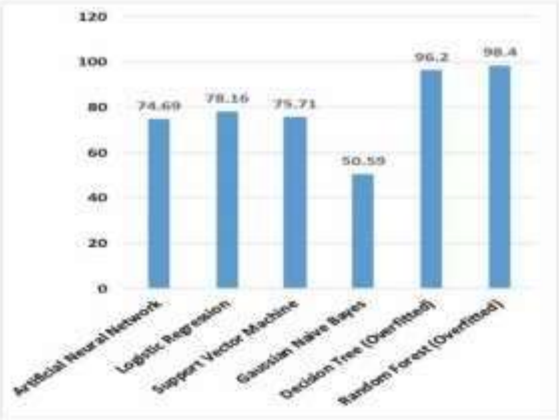


Fig. 6. Comparision of Results

VI. CONCLUSION

College students are suffering from many mental health problems including psychological distress, somatization, obsessiveness, interpersonal sensitivity, depression, anxiety, hostility, fear, paranoia and psychosis, which can have a number of negative effects on them. ML is about predicting the future based on past data. We use ML techniques to predict student stress. There is also a need of a well developed real time application so that students can use the application to determine their levels of stress.

REFERENCES

[1] Paper Title: Predicting the stress level of students using Supervised Machine Learning and Artificial Neural Network (ANN), Authors: Suraj Arya, Anju, Nor Azuana Ramli, Year: 2024

[2] Paper Title: Stress Detection in College Students Using Machine Learning Algorithm, Authors: Ms. Ancy Paul, Ms. Resija P R, Year: 2024

[3] Paper Title: Predictive Analysis of Student Stress Level using Machine Learning, Authors: Dr. Anbarasi M, Sethu Thakkilapati , Veeragandham Rajeev Vas, Sarabu Venkata Bharath Viswath , Year: 2023

[4] Paper Title: Stress Prediction in Indian School Students Using Machine Learning, Authors: Aanchal Bisht, Shreya Vashisth, Muskan Gupta, Ena Jain, Year: 2022

[5] Galabawa, J. Lwaitama,A Community Secondary schools: How long is their Journey to Quality Education? Paper presented at the University of Dares Salaam, Tanzania,(2018).

[6] Adler, A.K., and Wahl, O.F. Children's beliefs about people labelled mentally ill. American Journal of Orthopsychiatry,(2018)

[7] Brockington, I.F.; Hall, P.; Levings, J.Murphy, C.The community's tolerance of the mentally ill. British Journal of Psychiatry, 162:pp93-99, (2017).

[8] Corrigan, P.W.; River, L.; Lundin, R.K.; UphoffWasowski, K.; Campion, J.; Mathisen, J.; Goldstein, M.A.Stigmatizing attributions about mental illness. Journal of Community Psychology, (2010).

[9] Jiang, H. B, and Yang, D L, "Application Research on Fast Discovery of Association Rules Based on Air Transportation," International Conference on Service Systems and Service Management, pp. 1-6, 2007. 723

[10] Liu G, Jiang H, Geng R, and Li H, "Application of multidimensional association rules in personal financial services," International Conference on Computer Design and Applications , vol. 26, pp. 3877-3879, 2010.

[11] Qi W, Yan J, Huang S, Guo L, and Lu R, "The application of association rule mining in college students′mental health assessment system. Journal of Hunan University of Technology," vol. 6, pp. 94-99, June 2013.

[12] Meng Q, and Sha J, "Tree-based frequent itemsets mining for analysis of life-satisfaction and loneliness of retired athletes," Cluster Computing, vol. 2, pp. 1-9, May 2017.

[13] Huang S. C, Zhong J D, and Wen-Jua Q. I, "Statistical analysis and association rule mining of application in college students' mental health," Journal of Anyang Institute of Technology, pp. 108-111, 2014.

[14] Long Z, Feng H, Xue D, and Xiangjun D, "Positive and Negative Association Rules Mining for Mental Health Analysis of College Students," Journal of Mathematics Science and Technology Education, vol. 13, pp. 5577-5587, 2017.

[15] Herawan T, Vitasari P, and Abdullah Z, "Mining Interesting Association Rules of Students Suffering Study Anxieties Using SLP-Growth Algorithm," IGI Global, pp. 24-41, 2012.