Balancing the Scales: Employment Status, Educational Background, and Machine Learning in Predicting Student Anxiety

Ashvini Alashetty[1*], Saliha Bathool[2*], Jagdish Chandra Patni[3]

[1]School of CS & IT , JAIN (Deemed to be) University, Bangalore, India.

[2*,3]School of Design and Engineering, Alliance University, Bangalore, India.

[1*]ashwinialashetty@gmail.com
[2]*Saliha.bathool@alliance.edu.in
[3]Jagdish.patni@alliance.edu.in

Abstract. The issue of student anxiety remains an essential area of concern in academia and several reasons, like employment and educational levels, fuel it. This study gives a fresh perspective on how these considerations contribute to increasing or decreasing the levels of anxiety and implementing machine learning models for making predictive decisions involving anxiety. Frequentist hypothesis-testing statistical procedures with ANOVA and correlation tests showed the means differed by employment status shoulder to shoulder to ($F = 51.302$, $p = 5.83e\text{-}33$) and level of education ($F = 15.633$ $p = 3.82e\text{-}10$) level. The existing literature outline in this research presented the diversity of machines learning models in use extending to Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Decision Tree for increasing both prediction accuracy and efficiency. Out of these statistical analysis tools used, logistic regression was found performing better at 82% prediction who forecasted psychotic symptoms accuracy among the sample while random forest was quite equal on all factors in where outcoming performance of all metrics were tested. Such findings support the hypothesis of machine learning integration into school intervention practices for students' mental health, suggesting new opportunities for prevention and care.

Keyword: Anxiety, Statistics, Machine Learning, Random Forest, Logistic Regression.

1. Introduction

Data and computational power, through Machine learning technique, find and help the activities faster and better, especially in information-embedded workplaces. However, there are issues in that there are usually opaque models with complex black-box making it quite hard to interpret as well as an issue in transparency [1]-[4]. Physically, the transformation is at the height of the importance of the rapid change that the relevance of technology possesses. Particularly, student well-being and mental health have become central issues. Increased student size becomes a problem of anxiety; it has become a significant worrying mode over the last few years [5]- [6]. It has been observed that students experience anxiety in different forms: for instance, test anxiety, social anxiety or performance anxiety, and this in turn leads to poor academic outcomes, lack of interest in

S. K. Swarnkar et al. (eds.), *Proceedings of International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICSHIT 2024)*, Advances in Intelligent Systems Research 190,
https://doi.org/10.2991/978-94-6463-704-5_4

academic material and in general affects the education standards [7]- [8]. Tackling these issues entails an in-depth comprehension of the complex factors both inside and outside the classroom that students are under in relation to the academic demands. Every traditional methodology of measuring students' anxiety has got its own relevant barriers, for example a questionnaire or even performance in digital media analysis. There has been an inclining tendency in more advanced and complete ways of predicting student anxiety, so-called as whole person concept. This may also incorporate biometrics such as HRV (heart rate variability), SC (skin conductance) or even eye tracking methods, which are usually absent in the regular ones. Therefore, these indicators need to be employed to enhance the predictive scope on the physical appearance of anxiety in college students. Specific reasons underlie this apparent focus on the study of students' activities on online and face-to-face learning sites-lecturers and peers. This is why such research themes as peer dynamics, teacher's influence, support networks will facilitate further compassion toward a student's anxiety [9]. Educators obtain better results using this combination of machine learning techniques with active learning approaches: educational outcomes are achieved with minimal anxiety for the student. These methods allow to take preventive measures through early recognition of patterns and signs of Anxiety. Using machine learning tools where dispersion measures spanning different attributes such as academic performance level and participation rates, biological measures like heart and skin conductance are correlated with student performance, participation and other related measures. These also give opportunities for the analysts, who notice alarming signs for a 'nervous breakdown,' to act well in advance before such states develop. As a result, the incorporation of active learning principles will promote proper usage of response actions based on the indicator being targeted. It also brings more active participation from the learners which leads to better outcomes as well. For example, those who are at risk can benefit from additional measures and seem to benefit from real time interventions that could prevent or lessen stressful stimuli. As the focal point of our investigation, the data show that the use of machine learning techniques in conjunction with the biometric information and active pedagogical techniques helps to predict and control the anxiety of students. Among the data analysis techniques, we utilized such machine learning methods as Logistic Regression and Random Forest. This model has been found to have a decent performance, which is why it can be used for real life situations in levels health interventions. Our outcomes suggest that such intervention models used in these methods enhance students' engagement and academic achievement.

By incorporating machine learning with active learning, we will understand students' anxiety responses and have an individualized response to this condition, therefore creating a healthy and conducive academic environment for the healthy life of students and success at school. Emphasizing the wellbeing of learners amounts to addressing educational outcomes for the learners within contemporary schooling configurations. With a reduction in anxiety levels among learners, they become resilient and more engaging with equal opportunities in academics. This way, education transforms from a hostile space to a nurturing one where students are valued. On this matter, academic achievement and fulfilment of education becomes realistic when the levels of anxiety among students have been improved.

2. Literature Review

Acknowledging student anxiety as a critical issue in educational environments has spurred the creation of predictive models. Integrating active learning methods with machine learning algorithms presents a hopeful avenue for pinpointing and assisting vulnerable students. This review delves into the progression of this interdisciplinary domain. Initially, research emphasized identifying the factors that contribute to student anxiety. Mental health issues related to college students have always been complex to diagnose and have needed a lot of time. However, recent research work has adopted five machine learning algorithms such as logistic regression, KNN classification, decision trees, random forest, and stacking to classify these fast with high accuracy. For instance, stacking successfully found a predicted accuracy of 82.20 in one of the conducted types of research. Stress is closely associated with the deterioration of teenagers' and youth's mental state around the world and is characteristic of the period of transition from adolescence to adulthood. Using a data visualization method and random forest regressor algorithm, a study done in India established fifteen factors that cause stress among students out of the twenty-five and has R Square of 0.8042. In addition, another study tried to identify factors that might help to predict suicidal ideation and behavior among college students using settings from the French i-Share project. They adopted a risk algorithm based on random forests model with 70 candidate predictors Direct comparison with the previous studies of the current study is not possible as those used different kits and age groups The predictive accuracy of the model had an AUC of 0.8 sensitivity of 79% for girls, 81% for boys, and positive predictive Value for girls was 40%, and for boys, 36%. Four key predictors were identified: 12-month suicidal ideation, trait anxiety, depressive symptoms and self-esteem. Taken together, these papers underscore the need for sound diagnostic approaches of mental health among college students and the reliability of machine learning approaches to predicting multiple facets of mental health problems. Student performance prediction by applying data mining techniques to facilitate academic achievements. It seeks to provide useful insights on how these tools can best be implemented by educators in institutions of higher learning. It takes a systematic approach discussing such important issues as how to define student attainment, what are appropriate variables for selecting learners, and what machine learning models will be suitable. This is intended to make data mining more friendly to teachers so that they can use it in their teaching maximally [15]. Recently an e-learning system model was proposed for personalizing learning content based on students' emotions. Portable EEG headsets were employed to record brain electrical signals and KNN was utilized for real-time emotion recognition of the learner's state of mind. Additionally, through reinforcement learning it tests for emotions and recommends specific contents that ensure positive emotional states remain intact. The effectiveness of an e-learning system is usually measured in two aspects: more on the side of the system and other more on the side of the students' performance, their level of engagement and satisfaction, their learning outcomes. A study conducted on a sample of undergraduate computer science students learning at the University of Nottingham Ningbo China, in the academic year 2020-21 showed that the KNN algorithm was able to correctly identify the different emotional states to 74.3%, precision coming in at 70.8%, while recall stood at 69.3%. In addition, the analyses of t-Test clearly indicated that the proposed e-learning system generated higher level of satisfaction among students compared with the earlier traditional e-learning systems with statistical significance of p <

0.05. But then again within the context if learning and involvement the system made no difference. A study conducted in Peru to know the level of anxiety among the university students at the Universidad Nacional Mayor de San Marcos where machine learning based classification was used. In this paper, the model based on using RF, Naive Bayes, and SVM for the algorithms achieved 85% accuracy and F1=0.78 with average prediction time = 0.00100159s. Accomplishments to the methodology including data augmentation by Adversarial Generative Network and hyperparameters tuning using Grid Search and Random Search contributed to the improved performance of the developed model. Also, continued research is being conducted to find factors related to social processes belong to traditional and computer supported classrooms such as students' interaction each other, their interactions with studied subject and instructors, ways of getting help and assistances all these aspects are considered to define the term of anxiety. Current works use mobile data for identifying mental disorders such as depression, anxiety, and stress, but the relations between symptoms are not considered, nor are the time dependencies in the data. This paper therefore enriches a real-world parameter-sharing model for the detection of these disorders using 'sequential features' from mobile devices, leveraging on multi-task learning. The model yields an average accuracy of 0.78 and AUC of 0.78, which outperforms single-task models and traditional machine learning approaches, indicating the appropriateness of the multi-task learning approach in modelling severity of mental disorder symptoms.

Machine Learning is the approach to find a solution for complex datasets. Anxiety disorder is challenging to diagnose due to its various symptoms, leading to delayed treatment. Researchers use non-invasive bio signals like EEG, ECG, EDA, and RSP with machine learning to detect anxiety patterns. Studies show promising accuracies (55%-98%) using different bio signals, with EEG performing well, and EDA, RSP, and heart rate providing the most accurate results [18]. Mental health disorders like anxiety and depression have increased during COVID-19, observed through social media. Machine learning models were used to detect these disorders by analyzing users' online behavior and language patterns. Studies utilized data from various social media platforms to develop predictive models for identifying anxiety and depression, especially during pandemics [19]-[21].

3. Methodology

The methodology for analyzing anxiety levels in students using the given dataset involved a comprehensive approach encompassing data preprocessing, exploratory data analysis, statistical tests, and machine learning model implementation. Systematic research has been carried out to find the anxiety level in students. This dataset involved required information for finding the anxiety from the students. Our main goal is to find and understand the factors of the anxiety level among the students. The critical work in is to clean the data, eliminate the unnecessary columns or features, removing outliers, addressing missing information and feature scaling to make everything in an appropriate format or scale. Let's see step by step process which we have conducted in our research.

3.1. Data Description

Student Anxiety data collected from the Kaggle website, where the data size is 13464 rows and 55 features. This dataset indeed serves as a rich source of information for understanding student anxiety in depth. Overall, this dataset serves as a valuable resource for stakeholders invested in addressing student anxiety, offering insights that can inform research, practice, and policy initiatives aimed at promoting student well-being and success in educational settings [22].

3.2. Exploratory Data Analysis

EDA helps in understanding the basic structure and nature of the data, including the types of variables (e.g., numerical, categorical), their distributions, and summary statistics. This foundational understanding is essential before performing any further analysis or making inferences. This dataset contains summary statistics of various psychological and behavioral measures. The key variables include Generalized Anxiety Disorder (GAD) items, Satisfaction with Life (SWL) items, Social Phobia Inventory (SPIN) items, hours spent on activities, streams, narcissism scores, and age [23].

From Figure 1, we can observe that the students' anxiety levels vary significantly with their employment status. Specifically, the anxiety level is much higher during periods of unemployment, particularly between jobs and at the university level. In contrast, the anxiety levels at the school level and university level are similar. When students are employed, their anxiety levels are notably lower compared to when they are unemployed.

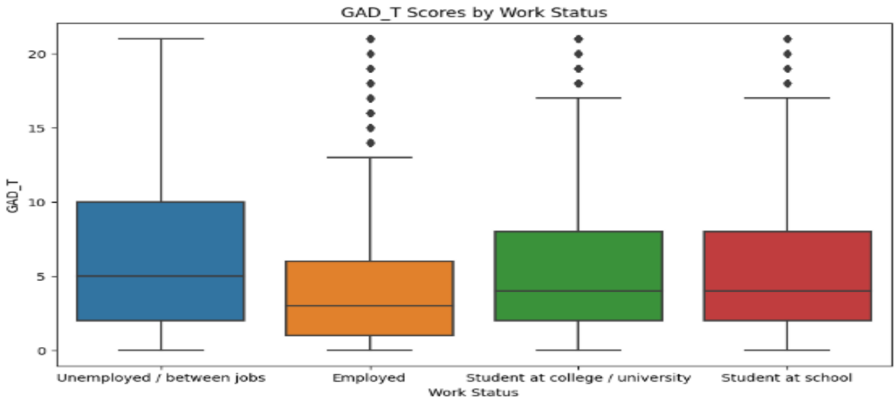

Fig 1. GAD_T Scores Across Different Employment Status

Figure 2 shows that young adults (ages 18-25) exhibit a wide range of GAD_T scores with numerous outliers, suggesting significant variability in anxiety levels. In contrast, individuals in their mid-twenties to early thirties (ages 26-33) show a narrower IQR, indicating less variability but still a presence of outliers. As we move to mid-thirties and late thirties (ages 34-39), the range of scores becomes more consistent, though outliers remain.

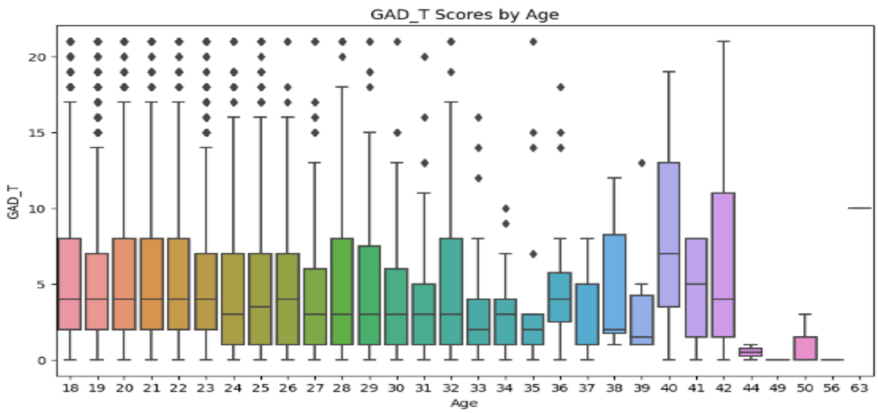

Fig 2. GAD_T Scores by Age

Notably, the forties age group (ages 40-42) displays increased variability again, with a broad range of scores and several outliers. Beyond this, ages 44-49 show smaller IQRs and fewer outliers, suggesting more consistent and lower anxiety levels. Individuals aged 50 and above exhibit very low variability, indicating very consistent and generally lower anxiety levels. This data highlights the shifting patterns of anxiety levels across different age groups, with younger adults experiencing more fluctuation and older adults generally showing steadier and lower anxiety levels.

The overall levels of generalized anxiety disorder and satisfaction with life are moderate, suggesting that while respondents do experience anxiety, it is not extremely severe, and their life satisfaction is reasonably high. The dataset predominantly consists of young adults aged 18-22, which could influence the interpretation of anxiety, life satisfaction, and social phobia levels due to the developmental stage and life circumstances typical of this age group. Variables such as 'Hours' and 'streams' show high variability and extreme values, indicating that there may be outliers or a wide range of engagement levels among respondents. As shown in table 1, certain items in the GAD and SPIN scales (e.g., GAD1, GAD3, SPIN11) have higher mean scores, indicating specific areas where respondents might be experiencing more significant symptoms.

Table 1. Summary Statistics of Psychological and Behavioral Measures of Anxiety

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
S. No.	13464.	7096.83	4114.47	1.000	3532.75	7087.50	10654.2	14250.
Timestamp	13464.	42054.8	0.27294	42052.	42054.7	42054.8	42054.9	42058.
GAD1	13464.	0.86096	0.92654	0.0000	0.00000	1.00000	1.00000	3.000
GAD2	13464.	0.67335	0.91572	0.0000	0.00000	0.00000	1.00000	3.000
GAD3	13464.	0.96576	0.98277	0.0000	0.00000	1.00000	2.00000	3.000
GAD4	13464.	0.72407	0.92197	0.0000	0.00000	0.00000	1.00000	3.000

GAD5	13464.	0.48804	0.83701	0.0000	0.00000	0.00000	1.00000	3.000
GAD6	13464.	0.91102	0.93116	0.0000	0.00000	1.00000	1.00000	3.000
GAD7	13464.	0.58875	0.89440	0.0000	0.00000	0.00000	1.00000	3.000
SWL1	13464.	3.72044	1.73626	1.0000	2.00000	4.0000	5.00000	7.000
SWL2	13464.	4.60205	1.69627	1.0000	3.00000	5.00000	6.00000	7.000
SW3	13464.	4.34544	1.80943	1.0000	3.00000	5.00000	6.00000	7.000
SWL4	13464.	3.76203	1.81813	1.0000	2.00000	4.00000	5.00000	7.000
SWL5	13464.	3.35888	1.91631	1.0000	2.00000	3.00000	5.00000	7.000
Hours	13434.	22.2473	70.2845	0.0000	12.0000	20.0000	28.0000	8.000
Narcissism	13441.	2.02767	1.06184	1.0000	1.00000	2.00000	3.00000	5.000
Age	13464.	20.9304	3.30089	18.000	18.0000	20.0000	22.0000	63.00
GAD_T	13464.	5.21197	4.71326	0.0000	2.00000	4.00000	8.00000	21.00
SWL_T	13464.	19.7888	7.22924	5.0000	14.0000	20.0000	26.0000	35.00
SPIN_T	12814.	19.8485	13.4674	0.0000	9.00000	17.000	28.0000	68.00

Initially, we loaded the dataset and inspected its structure to understand the variables and data types. We performed data cleaning by handling missing values and encoding categorical variables where necessary.
To explore the relationship between various factors and anxiety levels, we conducted a series of statistical tests. We began with the Shapiro-Wilk test to check the normality of the anxiety score distributions. Given the non-normality of the data, we applied the Kruskal-Walli's test and Mann-Whitney U test for non-parametric comparisons across different demographic groups, such as gender and age categories. These tests helped identify significant differences in anxiety levels between groups.

Insights into how different factors relate to anxiety levels were obtained from the results of these analyses as shown in table 2. ANOVA conducted on anxiety levels by work status shows dramatic difference in mean scores which means that whether a person is employed or not has significant impact on their level of anxiety. As a result, different job security for instance could be one reason that people in different employment statuses may experience varying amounts of anxiety.

Table 2. Statistical Analysis

Analysis	Statistic	Value	Degrees of Freedom	Significance (p-value)	Interpretation
ANOVA: Anxiety by work Status	F	51.302	3.134	5.83e-33	Significant difference in anxiety by work status
Correlation: Narcissism & GAD	Correlation ®	0.08	-	-	Weak Positive correlation
Correlation: Streaming & GAD	Correlation ®	0.00	-	-	No correlation
ANOVA: Anxiety by Degree	F	15.633	3.1183	3.82e-10	Significant difference in anxiety by educational degree

t-Test	t-Statistics	-14.148	-	4.06e-45	Significant difference between two groups
Normality Test: Group1	Shapiro-Wilk	-	-	0.0	Distribution is not normal
Leven's Test: Equality of Var	-	-	-	1022e-17	Variances are not equal

Among correlations between GAD and narcissism, a weak positive correlation was found indicating that those with high narcissism have slightly increased symptoms of GAD. However, because the strength of this relationship was not strong enough, it means that narcissism alone cannot predict the symptoms of GAD well enough. In contrast, no significant relationship is found between streaming hours and GAD symptoms, indicating that the amount of time spent streaming content does not impact levels of anxiety. This suggests that leisure activities such as streaming may not directly influence mental health outcomes related to anxiety.

ANOVA: The F-statistic value of 51.302, calculated using the ratio of the mean square between groups (MSB) to the mean square within groups (MSW), indicates a substantial variance between the work status groups relative to the variance within the groups. The degrees of freedom for the between-groups variation (df betweendf_{between}df between) is 3, which corresponds to the number of work status groups minus one. The degrees of freedom for the within-groups variation (df within df_{within}df within) is 134, reflecting the total number of observations minus the number of groups. The significance level (p-value) obtained from the F-distribution is 5.83e-33, which is exceptionally small. This p-value signifies that the likelihood of observing such a large F-statistic under the null hypothesis is virtually zero.

T-test: The t-test conducted shows a significant difference between two groups being compared, with the negative t-statistic indicating that the first group has a lower mean anxiety level than the second group. This suggests that there are distinct differences in anxiety levels between these two groups, which could have implications for understanding and addressing anxiety-related issues. The null hypothesis that the two groups have equal average anxiety levels is rejected because of the extremely low p-value. The t-test showed that there were statistically significant differences between the anxiety levels of the two groups. Normality and variance tests provide valuable information regarding distribution of data. Furthermore, Group 1's data fails to become normally distributed even after log transformation which means it should be cautious when interpreting statistical results. Logarithmic conversions also help reduce initial disparities in variances hence bringing into picture the possibility that extent of measurement could affect variance. U test affirms significant distinctions between these two groups as evidenced by t-test comparisons thereby reinforcing anxiety differences among them. It is apparent that such analysis can offer extensive understanding of the many interdependencies between distinct factors and anxiety, which might contribute to the formulation of remedies as well as support measures for diverse populations affected by this state. To begin with, we rescaled our data by taking logarithms of the same to meet up with ANOVA and t-test assumptions in order to compare anxiety levels across multiple demographic groups. Logarithm transformations were done on all variables except for one. Conversely, homogeneity of variance as a condition to ANOVA and t tests was indicated by Levene's test after an initial check on a normality assumption using Shapiro-Wilk test. Besides statistical analysis, some machine learning models were employed to predict anxiety levels by data set features.

Data splitting was done by using 80% of the data for the model construction and 20% for model evaluation. We used a range of techniques to predict the anxiety levels starting from basic logistic regression and using SVM, decision trees, random forest and other models. The performance of these models was measured using measures such as precision, recall, F-measure and accuracy. For example, the first logistic regression model we tested yielded 82% accuracy, though it reported different O/P for various classes in terms of precision and recall.

Consequently, this varied approach has offered a thorough exploration of students' stress levels identifying major demographic determinants of anxiety critically appraising various prediction techniques. This methodology therefore aims at gaining a better understanding of what data represents and how it could be useful for interventions that aim to reduce students' anxiety. Anxiety disorders can be detected through statistics or application of machine learning algorithms. Figure 3 describes an anxiety architecture using machine learning model that have the student anxiety data. Starting with collection of student data followed by exploratory data analysis (EDA) where we can observe and detect the patterns and find the errors if any. Any missing values are handled by level modelling using label encoding that will enhance the quality of data. As such, an updated dataset will now move into machine learning model selection stage where given problem requirements and nature of data determines what framework would be appropriate for use.

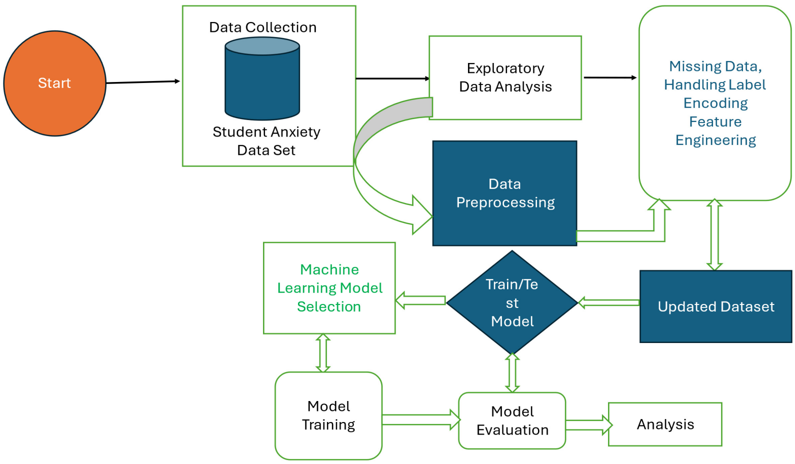

Fig 3. Student Anxiety Architecture

The chosen model is then trained on training data and tested on test data after this. Then an appropriate metric should determine which models work best and those which do not. A close analysis of evaluation results is carried out to interpret the outcome of a model that has been used herein including any potential issues that might arise because of this implementation. This can be rectified by going back to some previous steps like preprocessing or selecting another model, if necessary, since this process may become iterative until an optimal solution is reached through robust analysis on student anxiety. The flowchart as shown in figure 3, is for the process of identifying anxiety disorders via student anxiety dataset. Firstly, there was data collection stage where information on students' anxieties was gathered for future use in building models. It then proceeds to preprocessing which is also referred to as cleaning the data and it deals with issues like inconsistent records as well as missing values. After preprocessing, feature selection is carried out where only useful features are selected so that the performance of the model can be enhanced. For this process, t-test and ANOVA test among others statistical techniques have been applied to these selected attributes to determine whether they make sense or not for modeling purposes. In the case of statistics, machine learning model building comes next after statistical analysis where it has been built using selected features. Finally, this model will be evaluated in terms of its performance and accuracy with respect to identification of anxiety disorders. In summary, this evaluation helps us to know how effective the model is and identify areas that might need improvement. So basically, it can be noted that this flowchart gives an insight into various steps involved in constructing and evaluating a model aimed at recognition of anxiety disorders using students' information.

4. Result

The analysis of student anxiety using statistical methods and machine learning models yielded several significant findings. The primary goal was to identify factors influencing anxiety and evaluate different models' effectiveness in predicting anxiety levels. Class 0 indicates no or low anxiety levels. Class 1 indicates high anxiety levels. Overall, most models perform well in identifying Class 0 but struggle with Class 1. Logistic Regression, Random Forest, and K-Nearest Neighbors provide better balanced performance compared to SVMs and Decision Tree. Table 3 provides a comparative analysis of the precision, recall, F1-score, and accuracy metrics for each model, highlighting their strengths and weaknesses in predicting anxiety levels among students.

Table 3 Results of machine learning models

Machine Learning Model	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Accuracy	Comments
Logistic Regression	0.83	0.98	0.90	0.51	0.08	0.14	0.82	High Precision for Class 0
Random Forest	0.84	0.91	0.88	0.36	0.23	0.28	0.79	Moderate performance for class 1

Support Vector Machine (Linear)	0.82	1.00	0.90	0.33	0.01	0.02	0.82	Very low recall for class 1
SVM (RBF kernel, tuned)	0.82	1.00	0.90	0.33	0.01	0.02	0.82	Same as Linear SVM
K-Nearest Neighbors	0.84	0.93	0.88	0.36	0.18	0.24	0.80	Moderate recall for class 1
Decision Tree	0.84	0.88	0.86	0.32	0.25	0.28	0.77	Moderate performanc e

The below 5,6 &7 figures provided represent different performance metrics of several machine learning models used to predict anxiety levels in students. Here's an explanation of each graph:

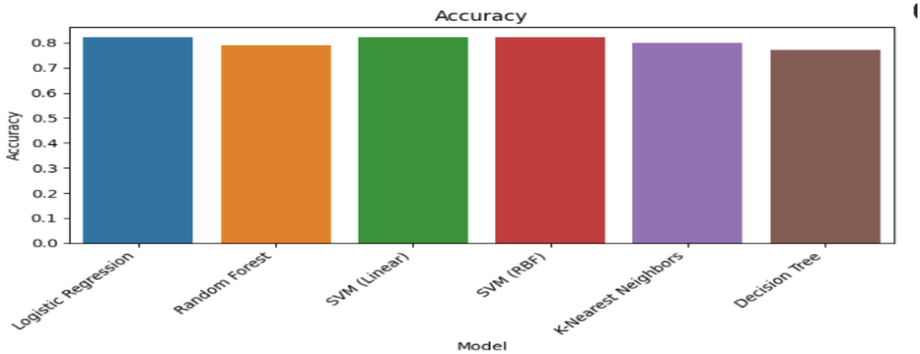

Fig 4. Models Accuracy

Accuracy measures the overall correctness of the model's predictions. It is the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. In shown in the figure 4 logistic regression method having the highest accuracy 0.82 as compared to other methods.

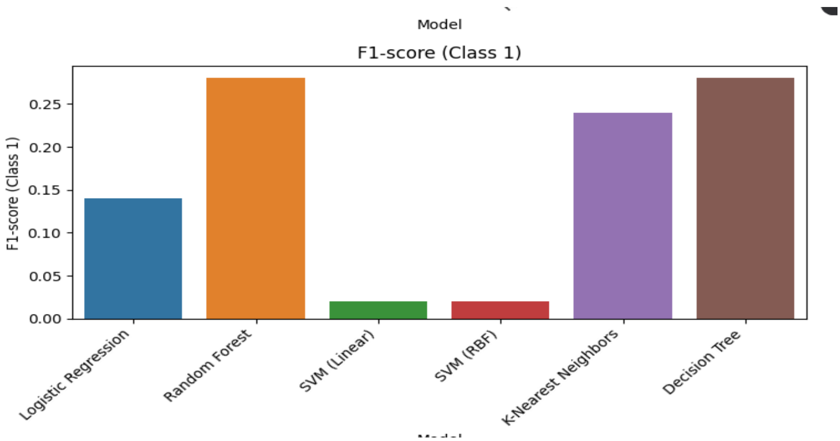

Fig 5. Model F1-score

The F1-score for Class 1 (students with anxiety) is the harmonic means of precision and recall, providing a balance between the two metrics. As shown in figure 5, random forest has the better F1 score as compared to other methods. Random Forest has the highest F1-score for Class 1, indicating it balances precision and recall effectively for predicting anxiety. Decision Tree and K-Nearest Neighbors also perform relatively well. Logistic Regression has a lower F1-score, while SVM (both linear and RBF) have the lowest F1-scores for Class 1, indicating these models struggle more with predicting the minority class (students with anxiety).

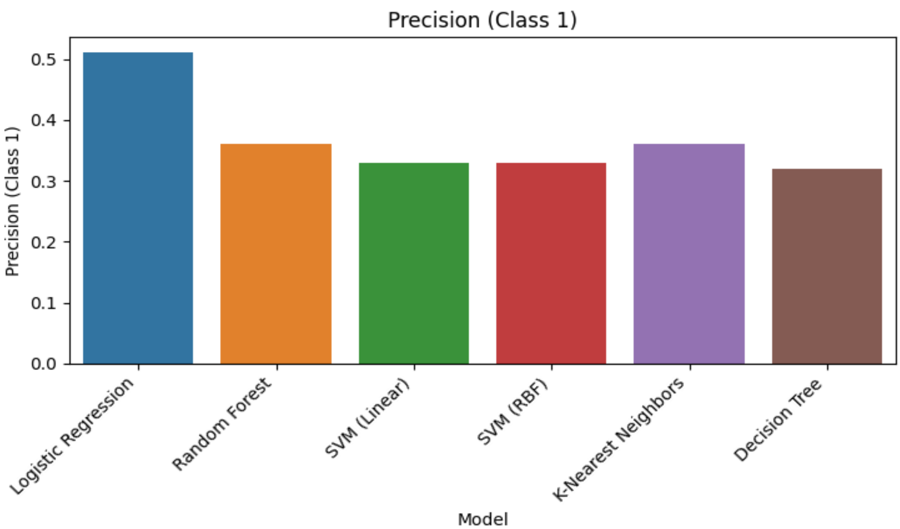

Fig 6. Precision (Class 1)

Precision for Class 1 measures the accuracy of the model's positive predictions; it is the ratio of true positives to the sum of true and false positives. As shown in figure 6, logistic Regression has the highest precision for Class 1, meaning when it predicts a student has anxiety, it is correct about half of the time. Random Forest, SVM (Linear), SVM (RBF), K-Nearest Neighbors, and Decision Tree all have lower precision scores, indicating a higher number of false positives in their predictions.

5. Conclusion

In our analysis of various machine learning models to predict anxiety levels among students, we evaluated Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree based on their precision, recall, F1-score, and overall accuracy. The Logistic Regression and Random Forest turned out to be the best with the Logistic Regression recording 82% overall accuracy in predicting low and high anxiety levels. However, Random Forest showed a balance between predicting low and high anxiety levels than other models and therefore it was better overall model. SVM models, while excelling at predicting low anxiety, struggled significantly with high anxiety predictions. KNN provided a reasonable balance but had moderate recall for high anxiety, while Decision Tree showed the lowest overall accuracy. Based on these results, we recommend using the Random Forest model for its balanced and robust performance in identifying students with varying levels of anxiety, making it well-suited for practical applications in addressing student mental health.

References

1. Biswas, T., Bhattacharya, D., Rudrapal, D., Roy, S. ,” Machine Learning for Forecasting Depression and Anxiety in University Students” In ICT: Applications and Social Interfaces, Lecture Notes in Networks and Systems, vol 908, Singapore, 2023, pp.83-95. https://doi.org/10.1007/978-981-97-0210-7_7
2. Pabreja, K., Singh, A., Singh, R., Agnihotri, R., Kaushik, S., Malhotra, T..,” Stress Prediction Model Using Machine Learning”, In Proceedings of International Conference on Artificial Intelligence and Applications. Advances in Intelligent Systems and Computing, vol 1164, Singapore, 2020, pp. 57-68. https://doi.org/10.1007/978-981-15-4992-2_6
3. Macalli, M., Navarro, M., Orri, M. et al., “A machine learning approach for predicting suicidal thoughts and behaviours among college students”, Sci Rep 11, 11363 ,2021. https://doi.org/10.1038/s41598-021-90728-z
4. Alam, A., Mohanty, A., “Predicting Students’ Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics” In Communication, Networks and Computing. CNC 2022. Communications in Computer and Information Science, vol 1893, Chem, 2023, pp.166-177, https://doi.org/10.1007/978-3-031-43140-1_15
5. Vera-Leon, B., Gozme-Avila, L., Guzmán-Monteza, Y., “ Classification Model for the Detection of Anxiety in University Students: A Case Study at UNMSM” In Good Practices and New Perspectives in Information Systems and Technologies, vol 989, 2024, Cham., pp.182-190, https://doi.org/10.1007/978-3-031-60227-6_16

6. Zhang, S., Tu, M., Yan, Y., Zhuang, Y., Ge, L., Wei, G., “Detecting Depression, Anxiety and Mental Stress in One Sequential Model with Multi-task Learning”, In HCI International 2022 – Late Breaking Papers: HCI for Health, Well-being, Universal Access and Healthy Aging. HCII, vol 13521, Cham., 2022, pp.191-200, https://doi.org/10.1007/978-3-031-17902-0_14
7. J. Southworth, K. Migliaccio, J. Glover, J. Glover, D. Reed, C. McCarty, J. Brendemuhl, and A. Thomas, ‘‘Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy,’’ Comput. Educ., Artif. Intell., vol. 4, Aug. 2023, Art. no. 100127, https://doi.org/10.1016/j.caeai.2023.100127
8. H. Halat, A. Soltani, R. Dalli, L. Alsarraj, and A. Malki, ‘‘Understanding and fostering mental health and well-being among University Faculty: A narrative review,’’ J. Clin. Med., vol. 12, no. 13, p. 4425, Jun. 2023, DOI: 10.3390/jcm12134425
9. X.-Q. Liu, Y.-X. Guo, and Y. Xu, ‘‘Risk factors and digital interventions for anxiety disorders in college students: Stakeholder perspectives,’’ World J. Clin. Cases, vol. 11, no. 7, pp. 1442–1457, Mar. 2023, doi: 10.12998/wjcc.v11.i7.1442.
10. K. Meeks, A. S. Peak, and A. Dreihaus, ‘‘Depression, anxiety, and stress among students, faculty, and staff,’’ J. Amer. College Health, vol. 71, no. 2, pp. 348–354, Feb. 2023, DOI: 10.1080/07448481.2021.1891913
11. R. Qasrawi, S. P. Vicuna Polo, D. A. Al-Halawa, S. Hallaq, and Z. Abdeen, ‘‘Assessment and prediction of depression and anxiety risk factors in schoolchildren: Machine learning techniques performance analysis,’’ JMIR Formative Res., vol. 6, no. 8, Aug. 2022, Art. no. e32736, DOI: 10.2196/32736
12. P. J. Bota, C. Wang, A. L. N. Fred, and H. P. Da Silva, ‘‘A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals,’’ IEEE Access, vol. 7, pp. 140990–141020, 2019, doi: 10.1109/ACCESS.2019.2944001
13. Pabreja, Kavita & Singh, Anubhuti & Singh, Rishabh & Agnihotri, Rishita & Kaushik, Shriam & Malhotra, Tanvi, “ Prediction of Stress Level on Indian Working Professionals Using Machine Learning”, International Journal of Human Capital and Information Technology Professionals, vol. 13. pp.1-26,2022, 10.4018/IJHCITP.297077.
14. Macalli, Melissa & Navarro, Marie & Orri, Massimiliano & Tournier, Marie & Thiébaut, Rodolphe & Côté, Sylvana & Tzourio, Christophe, “ Predicting suicidal thoughts and behaviors among college students: a machine learning approach”, 2021, 10.21203/rs.3.rs-227840/v1.
15. Syed Mustapha, S.M.F.D. “Predictive Analysis of Students’ Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods”, Appl. Syst. Innov. , vol 6, issue 86, 2023, https://doi.org/10.3390/asi6050086
16. Liu, X., Ardakani, S.P., A machine learning enabled affective E-learning system model”, Educ Inf Technol , vol. 27, pp. 9913–9934 , 2022, https://doi.org/10.1007/s10639-022-11010-x
17. Seo, K., Tang, J., Roll, I. et al., “The impact of artificial intelligence on learner–instructor interaction in online learning”, Int J Educ Technol High Educ, vol. 18, issue. 54, 2021, https://doi.org/10.1186/s41239-021-00292-9
18. Ancillon L, Elgendi M, Menon C., “Machine Learning for Anxiety Detection Using Biosignals: A Review” Diagnostics (Basel), vol. 12, issue 8, 2022, doi: 10.3390/diagnostics12081794. PMID: 35892505; PMCID: PMC9332282.
19. Verma, P., Sharma, K., Walia, G.S., “Depression Detection Among Social Media Users Using Machine Learning”, In International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1165, Singapore, 2020, pp. 865-874, https://doi.org/10.1007/978-981-15-5113-0_72
20. Liu D, Feng XL, Ahmed F, Shahid M, Guo J. ,”Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review”, JMIR Ment Health, Vol. 9, issue 3, 2021, doi: 10.2196/27244. PMID: 35230252; PMCID: PMC8924784.

21. Sofia, Arun Malik, Mohammad Shabaz, Evans Asenso, “Machine learning based model for detecting depression during Covid-19 crisis”, Scientific African, Vol. 20, 2023, e01716, ISSN 2468-2276, https://doi.org/10.1016/j.sciaf.2023.e01716.
22. Student Axexity Data, (Accessed on 28 28 April 2024), https://www.kaggle.com/datasets/petalme/student-anxiety-dataset