

Cluster Pattern Analysis of Students Stress using Machine Learning Algorithms with Feature Engineering

Delali Kwasi Dake*

Department of ICT Education, University of Education, Winneba, Winneba, Ghana

E-mail: dkdake@uew.edu.gh

ORCID iD: <https://orcid.org/0000-0002-6067-6980>

*Corresponding Author

Received: 12 May, 2025; Revised: 10 July, 2025; Accepted: 26 October, 2025; Published: 08 April, 2026

Abstract: Students attending school experience stress due to a variety of factors that can originate either on campus or from home. In managing stress and other related issues, educational institutions now have counselling units that operate as centres. Moreover, several institutions have designated academic counsellors in departments to address the increasing demand for counselling services due to an expanding student population. Timely detection and proactive counselling of stress among students can help avert dropouts, health issues, and other learner behaviours that are detrimental to academic work. This study proposes two approaches to facilitate the automation of student counselling for stress management. We first implemented the K-means algorithm and formed clusters using the elbow and the Silhouette methods. The clusters formed reveal three groups of students. The stressors significantly affected one group, making it vulnerable. The stressors moderately impacted another group, while the final group experiences minimal stress. In the second part of the study, we proposed a classification model to identify the cluster group of any new student. The results of the classification show superior performance for the Decision Tree algorithm with an accuracy of 97.64%. The improvement in the efficiency of the classification algorithms was attained through feature engineering using the Chi-square method.

Index Terms: K-means, Decision Tree, Random Forest, Student Stress Analysis, Machine Learning, Artificial Intelligence, Counselling, Education 4.0

1. Introduction

Stress among students in educational institutions has long been a concern since it affects learners' academic performance, mental health, and their overall quality of life [1, 2]. Learners face varying stress levels, including the persistent pressures of performing well academically, the monetary strain of paying fees and personal expenditures, relationship stress, and future-related stress of employment. The World Health Organisation (WHO) [3] defines stress as a natural human reaction that compels individuals to confront challenges and dangers in their lives. The United Nations Children's Fund (UNICEF) [4] describes stress as an unpleasant experience that occurs when we are overwhelmed, under pressure, or unable to cope with a situation. Stress has also been defined in literature as a perceived peril to one's emotional, mental, spiritual, and physical well-being, prompting a cascade of physiological responses and adaptations [5, 6, 7]. The data on student stress in educational settings are quite alarming. A survey by Transforming Education [8] reveals that about 80% of American college students identify financial, academic, and daily stress as the most difficult aspects of their school experience. According to a recent study conducted by Research.com [9], 61% of students at American universities attribute their low academic performance to stress. In that same report, 40% of college students in the United States feel inadequately rested five out of seven days a week. This aligns with global trends, as other studies across continents indicate a high level of stress among university students [10, 11, 12].

Counselling units have generally focused on mitigating stress among university students [13, 14]. The majority of educational institutions now have counselling centres that are decentralised to the departmental level. The challenge, however, is the persistent growth of student numbers across tertiary institutions globally. According to the World Bank [15], there will be over 222 million students enrolled in tertiary institutions worldwide in 2025, an increase from 100 million in 2000. The same report indicates that in Sub-Saharan Africa (SSA), enrolment figures have increased every two decades since the late 1970s. The conventional method of isolating stressed students for therapy sessions is

unproductive due to limited time constraints from a growing student population. Automating stress management for tertiary students will facilitate early diagnosis and mitigate severe acts, such as dropout, academic failure, and other adverse behaviours.

Machine learning (ML) demonstrates strong automation capabilities with big data, yielding classification percentages for precise detections across many domains. Recently, machine learning applications in education have shown a positive trend in alleviating educational challenges. Clustering or unsupervised learning is a ML technique that identifies groupings in an unlabelled dataset by increasing intra-cluster distances while decreasing inter-cluster distances [16, 17, 18]. In education, cluster analysis applications include personalised learning, collaborative learning, student counselling, curriculum development, student monitoring in a virtual learning environment, and sentiment analysis [19, 20]. In contrast, supervised learning, also known as classification, uses a labelled dataset for predictive modelling [21, 22]. Classification domain applications in education include academic performance prediction, learner retention prediction, student lesson engagement prediction, at-risk student prediction, dropout mitigation, and automatic grading [23, 24]. The diverse application of machine learning with educational data has established the foundation for the emergence of Education 4.0, which depends on technology and automation to revolutionise teaching and learning. The transformation of the educational landscape is essential, as pedagogical integration, instruction, and learning must be adapted to Industry 4.0.

The principal objective of this study is to cluster students according to their stress levels for proactive counselling. Clustering identifies similar stress types and levels among students, which is important for group counselling. The second objective is to develop a predictive model that automates future stress clusters of new students. The efficiency of predictive modelling stems from its ability to recognise learner stress and respond automatically. This effectively addresses the issue of a rising student population and the continued use of conventional approaches to stress management and counselling in tertiary institutions. In line with the objectives, we pose the following research questions:

- (1) What is the appropriate number of clusters formed based on the stress level dataset
- (2) To what extent does cluster membership varies with respect to distinct features and similarities
- (3) Which supervised learning algorithm has the highest performance metrics in clustering stress before and after feature engineering.

The subsequent sections of the paper are organised as follows: Section 2 examines pertinent material concerning stress levels using machine learning. Section 3 implements and analyses the methodological approach. Section 4 conducts a rigorous analysis of the cluster simulation and its outcomes. Section 5 analyses the findings in relation to the literature. In Section 6, we present the conclusion and delineate the study's limitations.

2. Review of Literature

The literature review is organised into two sections. The first section of the review investigates basic research on factors that cause stress among students and the levels of stress they face. The second section of the review examines the implementation of machine learning in stress clustering, management and classification.

2.1. Stress and Stressors – Basic Studies

Stressors can be categorised as family-related, university-related, and psychosocial. The family-related stressors are mainly financial issues, divorce, lack of support, and hereditary diseases [25, 26, 27]. The university-related stressors include examinations, accommodation, relationships, subject-specific instructors, grade points, and academic workload [11, 28, 29]. The psychosocial stressors include diet, sleep quality, loneliness, anxiety, self-esteem, and depression [27, 30, 31].

Several studies have demonstrated that university students encounter diverse stressors and varying levels of stress while on campus. The following are some basic studies on stress factor determination and levels among university students. Alkhawaldeh et al. [26] investigated the stressors, stress levels, and coping strategies of university students at the Sultan Qaboos University (SQU). Their findings indicate that 11.4% of students experience light stress, 75.1% have considerable stress, and 13.5% have severe stress. Furthermore, the data showed that the highest-ranked factors causing stress were chronic illness, exam period, low cumulative grade point average (CGPA), and living alone. Shehada et al. [32] conducted a study at the Faculty of Dentistry, Damascus University, to understand factors causing stress among dental students. Their findings indicate moderate stress levels among dental students, attributing the primary sources of stress to academic workload, insufficient dental equipment, and challenges in sourcing clinical case requests. Brett et al. [33] investigated the well-being among university students in the United Kingdom (UK) and identified the factors that have a direct correlation with it. Their principal findings demonstrate that residential students experience greater levels of stress compared to commuters. Moreover, loneliness, feeling of coherence, and resilience emerged as the most significant indicators of students' well-being. Hanna et al. [34] used descriptive statistics to analyse stressors among pharmacy students at Queen's University Belfast in the UK. Their findings indicate that female students experience

more stress than male students. In addition, the most common stressors were financial, job opportunities, and career choice. In a related study by Ngaboyishema et al. [35] on stress management at the Adventist University of Central Africa, female students experience more stress than male students. Furthermore, they discovered that students majoring in finance have the highest levels of stress. Pitt et al. [36] employed a mixed-methods approach to assess the stress levels of Australian university students. Their findings identified academic, economic, interpersonal, university transition, and social support as the primary stressors.

2.2. Stress and Stressors – Machine Learning

Implementing machine learning in learner stress management will allow for more efficient, early, and confidential diagnosis of students' stressor features and severity levels. Clustering will result in group therapy for students with comparable stressor characteristics, whilst classification will feed predictive models of stress levels and groups. In the ML review, we will assess indicators for algorithmic performance comparison, identify the most prevalent algorithm, and examine validation strategies. Furthermore, a concise discussion of the limitations of each study is presented.

The following are machine learning studies on perceived stress level (PSS) and feature engineering for identifying prominent factors that influence PSS. Disha Sharma et al. [37] analysed the performance of machine learning algorithms to detect stress levels among 200 undergraduate students for proactive counselling. The class label, PSS, consists of normal, moderate, and high stress categories. Random Forest (RF) emerged as the best-performing algorithm with an accuracy of 94.73%. The study evaluated the RF method to several techniques, including Bayes Net, Naïve Bayes (NB), Logistic Regression, J48, and multilayer perceptron (MP). Their investigation, however, did not identify the prevalent stressor and was confined to only classification.

Bhatia et al. [38] proposed a machine learning model to classify academic stress among 1320 students from colleges in UT Chandigarh. The PSS was categorised as low, medium, and high. Approximately 49.34% of the students experienced moderate stress. Four machine learning classifiers, including ML-KNNweb, classifier chains, binary relevance, and OneVsRest, were employed for performance assessment. Their research ranked the following stressor factors: homesickness, appearance, family pressure, future job uncertainty, concentration, health, study workload, examination, teacher and student conduct, and finance. The study workload, examination, and future job uncertainty were identified as the most influential factors in PSS categorisation. The study, however, did not disclose the top-performing classifiers. Furthermore, the classifier's design did not include any cross-validation technique.

Paul & Resija [39] investigated three types of PSS: acute, episodic, and chronic, among 954 Vimala College students using machine learning algorithms. The classifier was built using multiple machine learning methods, including random forest, logistic regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), linear regression, Adaboost, Hist gradient boosting, and decision tree (DT). The RF algorithm emerged as the superior classifier with an accuracy of 99%. The study, however, was limited to classification and did not identify the most significant feature influencing the PSS.

Rois et al. [5] proposed a ML model for stress detection among 355 students from twenty-eight different Bangladeshi universities. They employed the Boruta algorithm for feature engineering to identify prognostic features of PSS and applied DT, RF, SVM, and linear regression techniques in developing a prediction model. Their findings indicate that students' academic background, smoking status, sleep state, systolic and diastolic blood pressures, and pulse rate were significant factors in predicting PSS. The RF algorithm was identified as the most effective method, achieving an accuracy of 89.72%. The study, however, was limited to only classification.

Ahuja & Banga [40] assessed stress levels among 206 students at the Jaypee Institute of Information Technology via machine learning methods. The RF, NB, SVM, and KNN classifiers were compared and implemented using the k-fold cross-validation technique. The results showed that the SVM algorithm performed best, with an accuracy of 85.71%, followed by the RF method at 83.33%. However, the study focused solely on classification and did not investigate dominating features in determining PSS.

Firoz et al. [41] proposed a machine learning approach to assess mental stress among 1632 Bangladeshi undergraduates. The study employs features such as CGPA, sleep deprivation, relationship challenges, financial issues, food aversion, and substance abuse. In building the classifier, the RF, DT, KNN, MultinomialNB, and the SVM were evaluated for performance. The RF and DT classifiers achieved the highest validation and test scores of 99%. However, the study was limited to classification, and feature engineering was not considered in the final analysis.

Ding et al. [42] based in an ensemble machine learning model investigated stress among 630 students. The model was built with the student's heart rate, sleeping rate, eye movement rate, blood oxygen levels, limb movement rate, body temperature, respiration rate, and snoring range as features. The stress predictor label consists of the following: low, medium low, medium, medium high, and high. The proposed ensemble classifier comprises a combination of Random Forest (RF) and Gradient Boosting Machine (GBM). The combined model achieved an accuracy of 100%. Although the study employed classification, deep learning, and statistical techniques to determine the superiority of the hybrid model, clustering and feature engineering were not examined.

Gorabal et al. [43] proposed a machine learning model for stress detection by examining the efficacy of RF and DT algorithms in a cohort of 2,958 students. They identified the following attributes as stressors while constructing the model: social expectation, academic pressure, financial concerns, and personal issues. The DT emerged as the best performing algorithm with an accuracy of 99.46%. Nonetheless, clustering and feature engineering were not implemented in this investigation.

Filippis et al. [1] investigated stressors among 1100 students using social, academic, environmental, physiological, and psychological factors. They employed the RF approach to develop a predictive model for the dataset and prioritise the most critical stressors affecting the students. Their results identified self-esteem and sleep quality as critical predictors of stress levels in students. Their research, however, has drawbacks, such as the lack of a comparison of the RF algorithm to other classification methods. The RF algorithm achieved an accuracy of 88%

Table 1. Literature Summary – Machine Learning

Review	Classification	Clustering	Algorithms	Best Performing Algorithm	Major Limitation
Disha Sharma et al. [37]	✓	✗	RF, Bayes Net, NB, Logistic Regression, J48, and MP	RF, accuracy = 94.73%	(i) No feature selection (ii) Only classification
Bhatia et al. [38]	✓	✗	ML-KNNweb, Classifier chains, Binary relevance, and OneVsRest	✗	(i) No accuracy reported (ii) Only classification
Paul & Resija [39]	✓	✗	KNN, SVM, NB, Linear regression, Adaboost, Hist gradient boosting, and (DT)	RF, accuracy = 99%	(i) No feature selection (ii) Only classification
Rois et al. [5]	✓	✗	DT, RF, SVM, and Linear regression	RF, accuracy = 89.72%	(i) Only classification
Ahuja & Banga [40]	✓	✗	RF, NB, SVM, and KNN	SVM, accuracy = 85.71%	(i) No feature selection (ii) Only classification
Firoz et al. [41]	✓	✗	RF, DT, KNN, MultinomialNB, and the SVM	RF, DT, accuracy = 99%	(i) No feature selection (ii) Only classification
Ding et al. [42]	✓	✗	RF and GBM	Hybrid RF and GBM, accuracy = 100%	(i) No feature selection (ii) Only classification
Gorabal et al. [43]	✓	✗	RF and DT	DT, accuracy = 99.46%	(i) No feature selection (ii) Only classification
Filippis et al. [1]	✓	✗	RF	RF, accuracy = 88%	(i) Only classification

Table 1 demonstrates that clustering has not been applied to managing the incidence of stressors and stress levels among students. However, for effective group counselling with similar learner stressors, clustering is a mandatory machine learning technique to be applied. The proposed study will investigate the use of clustering in learner stressor management and build a classification model for future cluster group prediction.

3. Research Methodology

As depicted in Fig. 1, the study consists of two parts. The initial segment pertains to clustering, while the subsequent segment concerns classification.

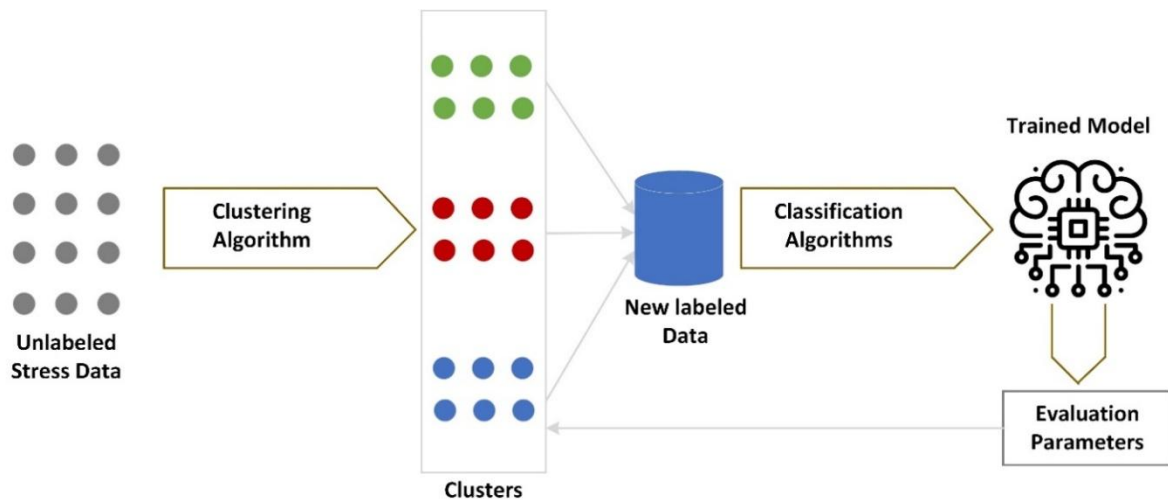


Fig. 1. Proposed methodology diagram

3.1. Unlabelled Stress Data

The dataset for this stressor research is acquired from GitHub (<https://github.com/Foysal440/Stress-analysis/blob/main/StressLevelDataset.csv>). The dataset consists of 20 stressors and 1100 student responses across various tertiary institutions. As shown in Table 2, the 20 stressors are grouped into psychological factors, physiological factors, environmental factors, academic factors, and social factors.

Table 2. Stress Dataset from Github

Category / Factors	Stressors
Psychological	anxiety level, self-esteem, mental health history, depression
Physiological	headache, blood pressure, sleep quality, breathing problems
Environmental	noise level, living conditions, safety, basic needs
Academic	academic performance, study load, teacher-student relationship, future career concerns
Social	social support, peer pressure, extracurricular activities, bullying

3.2. Clustering Algorithm

The dataset informs the selection of the clustering algorithm. K-means groups similar data points based on numeric data, K-modes based on categorical data, and K-prototypes based on both numeric and categorical data [44, 45]. The study's dataset is numeric, hence necessitating the application of the K-means algorithm. Although hierarchical clustering functions adequately with numeric datasets, it is more efficient for datasets characterised by a limited number of instances and attributes. Moreover, hierarchical clustering does not necessitate a pre-established number of clusters, complicating the identification of the best cluster count for certain types of data [46, 47]. In this study, the number of data instances and features are quite moderate, and based on the dataset, the number of clusters is envisaged ahead of time. These characteristics make the K-means algorithm appropriate for this research.

3.2.1. K-means Clustering

K-means is a distance-based or a centroid-based clustering algorithm that partitions an unlabelled dataset into similar groups. The K-means technique employs the cluster centroid to minimise the sum of squared Euclidean distances between the centroids and their respective cluster members. The primary objective of the K-means algorithm is to reduce the inter-cluster variance and widen the intra-cluster similarities. The K-means technique involves the initialisation of data into random clusters. The randomly designated clusters are subsequently allocated a centroid. Each data point converges towards a cluster centroid by iterative application of a distance metric. The clusters are continuously redistributed based on the similarity of data points until convergence is achieved, while minimising dissimilarity [48, 49, 50]. Mathematically, K-means partitions n instances of data into k clusters where $k < n$, and each data instance is assigned to the cluster with the closest mean. Equation 1 demonstrates the K-means algorithm's application of Euclidean distance to determine the separation between two data points.

$$Dist_{qp} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (1)$$

In Equation 1, q and p represent sample data points within an n -dimensional feature space, yielding a multivariate mean in Euclidean space. The K-means technique is effective with numerical datasets but underperforms with categorical and mixed data types. As shown in the process flow diagram in Fig. 2, the K-means uses a series of algorithmic steps to arrive at the number of centroids.

3.3. Classification

The classification component of the study will forecast future clusters of new students. This study will examine the machine learning efficacy of the Random Forest (RF) method, the Decision Tree (DT) algorithm, the Support Vector Machine (SVM), and the K-Nearest Neighbor (KNN). The algorithms were selected due to their superior performance in the literature

3.3.1. Random Forest (RF)

Random Forest is an ensemble machine learning algorithm. The RF algorithm creates a new dataset from an existing one by bootstrap aggregation. The RF classifier utilises a synthesis of decision trees to enhance the precision of its predictions [51, 52].

3.3.2. Decision Tree (DT)

The Decision Tree is a non-parametric supervised learning technique with a hierarchical structure made up of the root node, branches, internal nodes, and leaf nodes. To ascertain the feature that serves as the root node, methods such

as information gain, entropy, and Gini index are employed to evaluate the efficacy of a feature in classifying the provided data [53, 54].

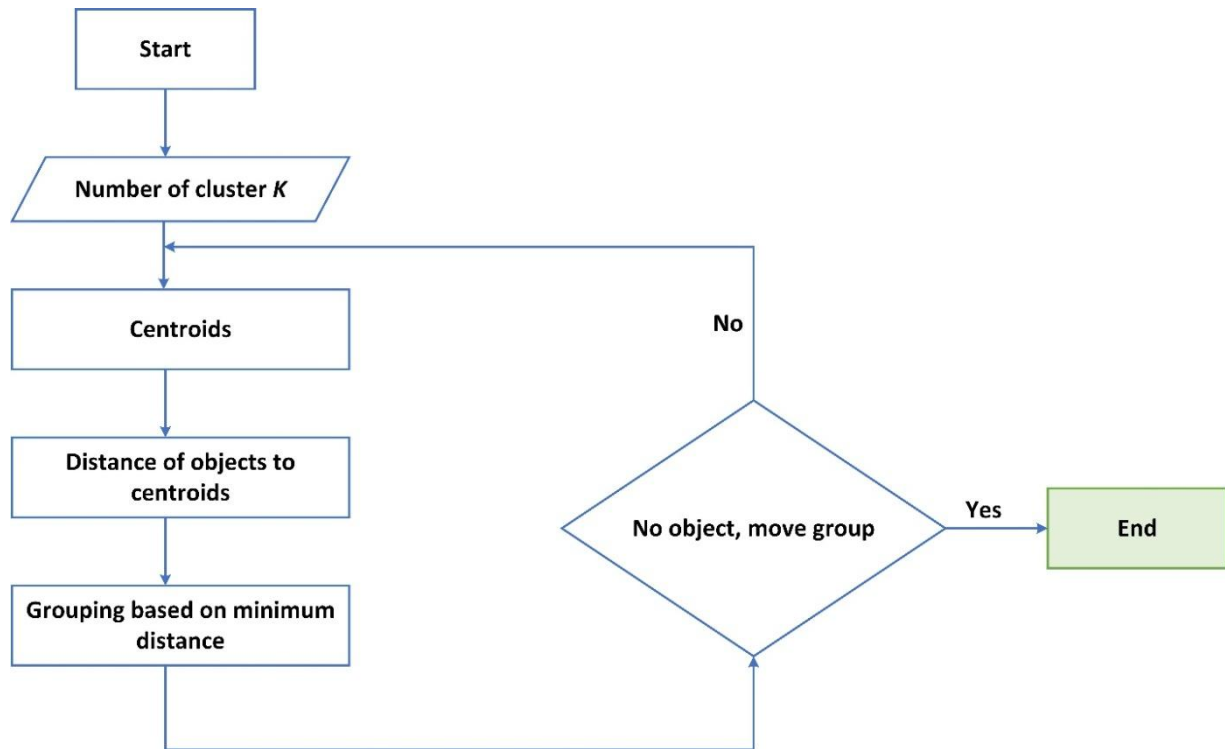


Fig. 2. Process flow diagram of the K-means algorithm

3.3.3. Support Vector Machine (SVM)

Support Vector Machines are binary classifiers that utilise vector representations of training cases to identify a hyperplane separating two classes. The SVM explicitly identifies the nearest point and delineates a line by vector subtraction [46], [48].

3.3.4. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor is an instance-based learning method that utilises proximity to evaluate a data point against a pre-existing dataset for predictive analysis. The value of K in the KNN algorithm is selected based on the input data. A noisy input dataset will require a higher selected value of K [56, 57]

3.4. Clustering Parameters

The primary clustering parameters investigated aim to determine the number of clusters using both the elbow method and the Silhouette method in order to arrive at the optimal cluster.

3.4.1. The Elbow method

The elbow function or method is utilised in clustering to ascertain the optimal number of clusters. The number of clusters k using the elbow method plateaus at a point where the Within-Cluster Sum of Squares (WCSS) values decreases as the number of clusters k increases [58].

3.4.2. The Silhouette method

The Silhouette method is a distance-based metric that illustrates the proximity of one cluster point to an adjacent cluster. It uses the Silhouette coefficient to determine the cohesion among clusters. The Silhouette coefficient varies from -1 to 1 , with a higher value indicating a model characterised by more cohesive clusters [59].

3.5. Classification Performance Metrics

The following classification metrics accuracy, precision, recall, F-measure, and the receiver operating characteristics curve (ROC-AUC), and the confusion matrix values were compared to select the best classifier. Equation 2 demonstrates that classification accuracy is directly proportional to the ratio of correct predictions. This metric quantifies the ratio of accurate forecasts to the overall number of predictions made.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

TN, TP, FN, and FP denote the number of true negatives, true positives, false negatives, and false positives, respectively.

The precision metric identifies true positives. Equation 3 determines precision by dividing the count of true positives by the total of true and false positives.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

The recall measure assesses the quantity of pertinent positive instances. Equation 4 computes recall by dividing the count of true positives by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The F-measure is a quantitative metric utilised to evaluate the efficacy of classifiers. Equation 5 illustrates that the F-measure is derived from the harmonic mean of precision and recall.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

The ROC curve evaluates a classification model's efficacy and highlights challenges in accurately categorising data across various thresholds. The ROC curve depicts the correlation between the True Positive Rate (TPR) and the False Positive Rate (FPR). The Area Under the Curve (AUC) quantifies the distinction between classes. Augmenting the AUC value of a classifier improves its ability to differentiate between specified classes.

4. Cluster Simulation and Results

In simulation, we used the Anaconda Python distribution from Continuum Analytics (Conda) along with SciPy, Pandas, NumPy, Matplotlib, and Scikit-learn. The presentation layer integrated Python source code and markdown text to construct a canvas in Jupyter Notebook.

4.1. Detecting the Number of Clusters

Research Question 1: What is the appropriate number of clusters formed based on the stress level dataset.

To respond to research question 1 and select the number of clusters, the study relied on the elbow and Silhouette methods. As shown in Fig. 3, for the elbow method, the Within-Cluster Sum of Squares (WCSS) values are important in selecting the optimal cluster. The point of inflexion, termed the elbow, has values that decrease as the number of clusters increases. The number of clusters, k selected using the elbow method is 3.

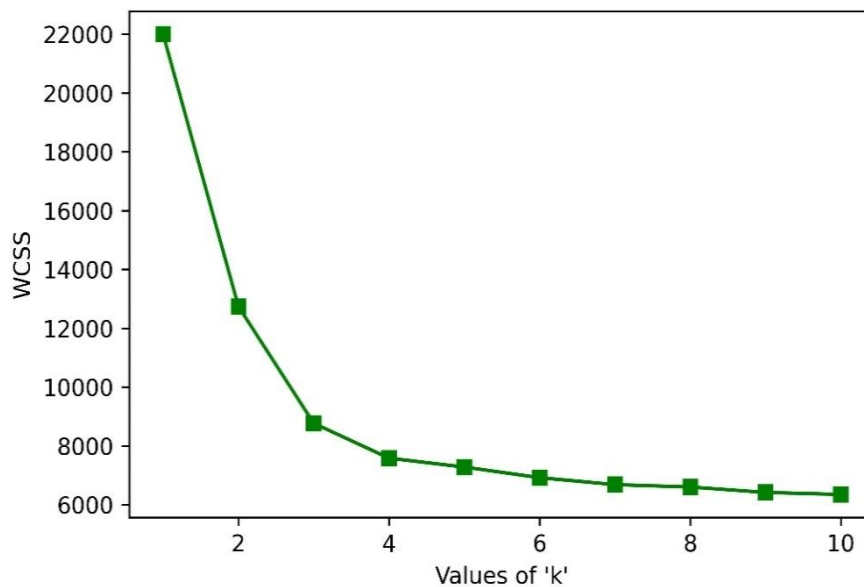


Fig. 3. Optimal cluster number using the elbow method

The Silhouette score in the Silhouette method for the number of clusters formed determines the cohesion among the clusters. Fig. 4 demonstrates that the silhouette score attains its maximum value at 5. The disparity in silhouette score values between clusters 3 and 4 and cluster 5 is, however, negligible. Selecting either cluster 3, 4, or 5 via the silhouette approach will yield relatively optimal cluster performance and cohesion. The number of clusters selected using the Silhouette method is 3. This corresponds to the chosen number of clusters using the elbow technique.

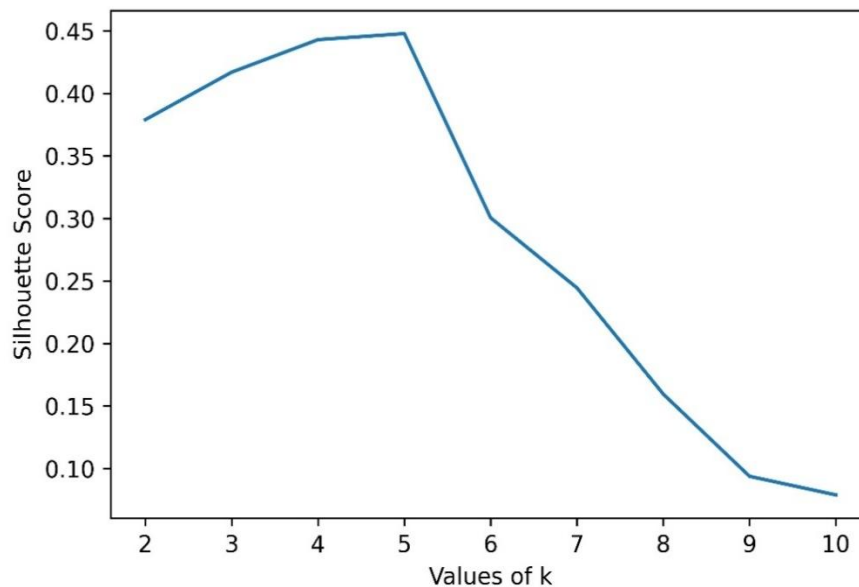


Fig. 4. Optimal cluster number using the Silhouette method

4.2. Cluster Member Distribution and Variation

Research Question 2: To what extent does cluster membership varies with respect to distinct features and similarities.

In response to research question 2, the cluster member variation analysis was performed in three stages. First, establish the number of cluster members; second, determine how each cluster member varies based on the features; and third, depict the cluster cohesiveness plot.

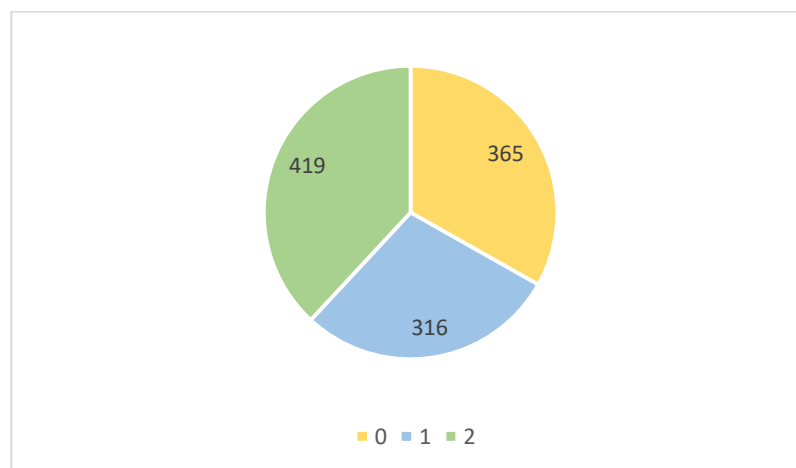


Fig. 5. Cluster members

As shown in Fig. 5, Clusters 0, 1, and 2 have different numbers of members. Cluster 2 has the highest membership of 419. Subsequently, Cluster 0 comprises 365 members, whereas Cluster 1 has the lowest number of members at 316. In percentage terms, Cluster 2 has 38.1%, greater than 33.2% of Cluster 0 and 28.7% of Cluster 1.

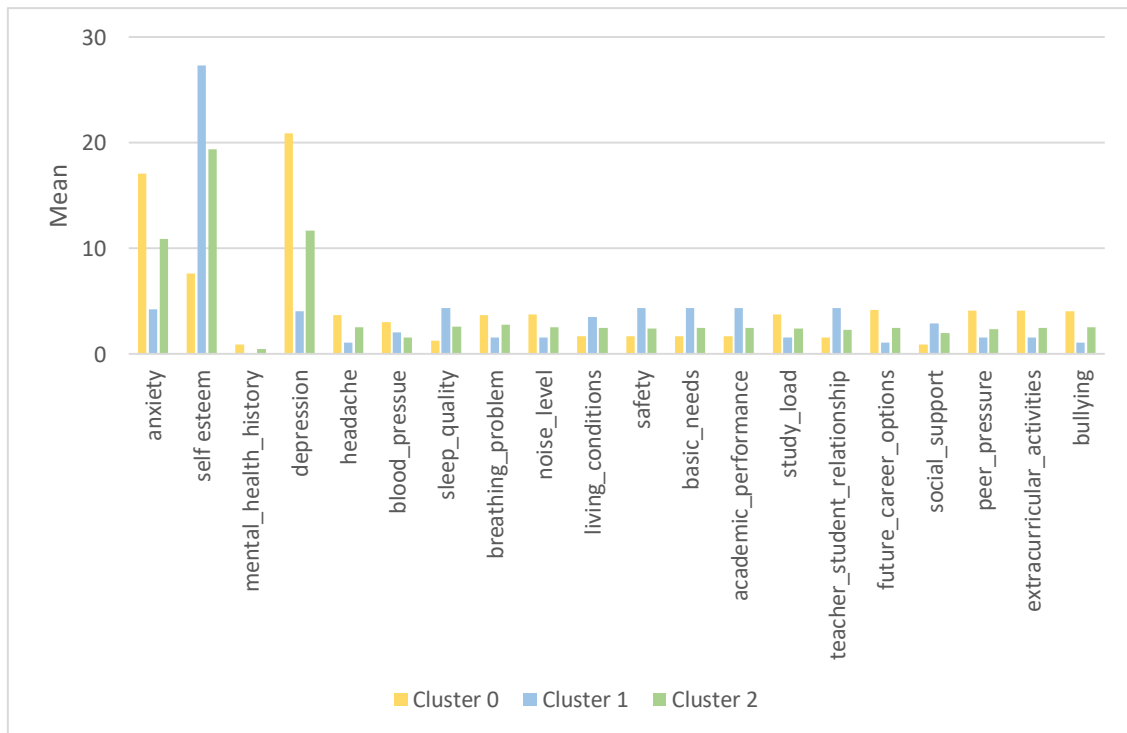


Fig. 6. Cluster Distribution with Features

Fig. 6 displays the cluster distribution in relation to the stressors. The mean value of the 20 stressors from the clusters formed and grouped under psychological, physiological, environmental, academic, and social categories are unique in understanding the cluster members. The stressors have less impact on Cluster 2 members. Despite the relatively high influence of peer pressure and frequent bullying on students in Cluster 2, they maintain relatively stable levels of anxiety, depression, and self-esteem. As a result, Cluster 2 members have the lowest blood pressure. Students in Cluster 2 do not require frequent counselling since the factors that contribute to stress among them are relatively low. Cluster 1 members have the highest academic scores, with the highest self-esteem, living condition, social support,

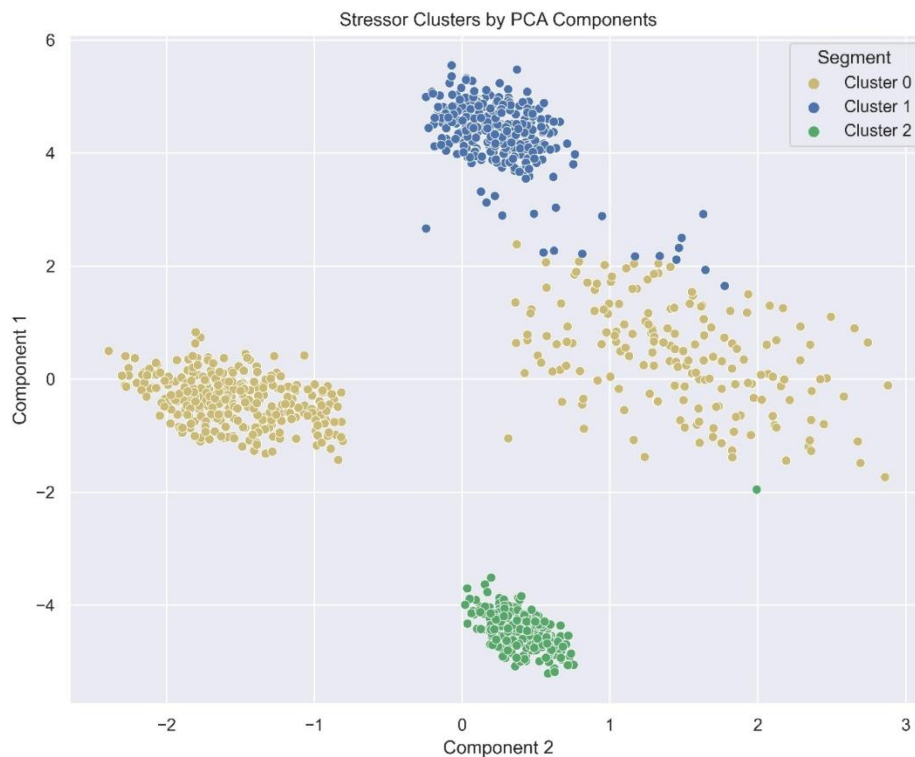


Fig. 7. Stressor Clusters using K-means with Principal Component Analysis (PCA)

basic needs, safety, and sleep quality. They do not participate in extracurricular activities; they are the least bullied with the lowest peer pressure. The blood pressure of Cluster 1 members, however, is moderately high. We can infer that Cluster 1 members are from financially stable homes and their family support systems are extremely high. This Cluster needs the least attention from the counselling unit compared to the other two clusters. Members of Cluster 0 exhibit the poorest academic performance alongside the most significant mental health issues, marked by elevated anxiety, diminished self-esteem, severe depression, frequent headaches, substandard housing conditions, inadequate teacher-student relationship, and a fragile social support network. Furthermore, individuals of Cluster 0 engage extensively in extracurricular activities; yet, they often encounter bullying and peer pressure. This cluster group exhibits highly elevated blood pressure and is significantly affected by noise. Members of Cluster 0, as determined by inference, belong to households with low income. These are at-risk learners who require regular counselling while on campus.

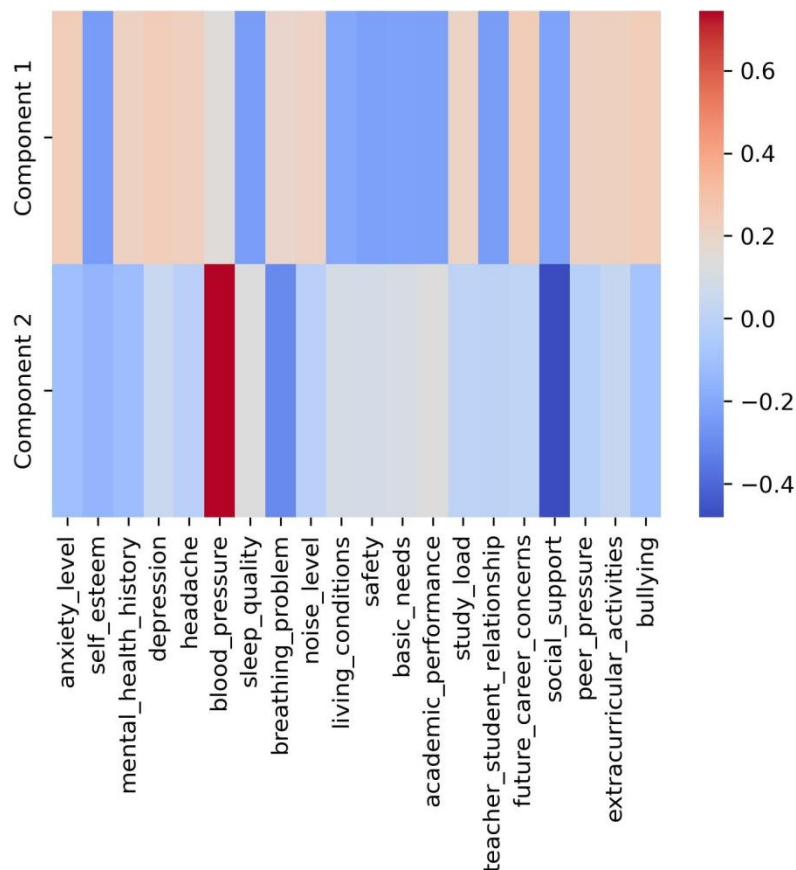


Fig. 8. Correlation heatmap between the stressors using PCA for cluster formation

In clustering, we employ Principal Component Analysis (PCA) as a dimensionality reduction method to identify pertinent features for enhanced cluster segmentation. Numerous studies indicate that K-means algorithms exhibit enhanced performance when PCA is utilised to diminish noise [60, 61]. The PCA in K-means as shown in Figure 7 visually separates the three clusters by reducing the inter-cluster variance and widening the intra-cluster similarities. Members of Cluster 0 are relatively dispersed, exhibiting some extreme values. This implies that individual counselling for Cluster 0 members will be more appropriate than group counselling. However, Cluster 2 members are cohesive, indicating that generalised group counselling will be appropriate. Cluster 1 members exhibit cohesiveness in their group with few outliers. A customised group therapy approach will significantly benefit this cluster's members.

The clusters formed in Figure 7 are influenced by some stressor features in the PCA. As illustrated in Figure 8, strongly correlated stressors are indicated by dark red colors while weakly correlated features are indicated by dark blue colours. The heatmap graph shows that blood pressure has a positive influence in cluster segmentation using the PCA while social support has the least influence in the cluster segmentation process.

4.3. Classification Results

Research Question 3: Which supervised learning algorithm has the highest performance metrics in clustering stress before and after feature engineering.

In response to research question 3, we evaluated the predictive capabilities of RF, DT, SVM, and KNN algorithms employing the tenfold cross-validation method. The tenfold cross-validation divides the dataset into ten subsets, utilising nine for training and one for testing; this procedure is performed ten times, employing new subsets for testing on each occasion. The cross-validation procedure is relevant to avoid problems of overfitting and underfitting of the

model. To identify the top-performing features that influence a student's cluster, the Chi-square feature selection method was used. To decrease computational cost, minimise data dimensionality, facilitate model training, and enhance model performance, the Chi-square feature engineering test is conducted to find the least relevant attributes that have negative, little, or no effect on the classifiers' performance. The following machine learning metrics, including accuracy, F-measure, ROC-AUC, and confusion matrix were evaluated to determine the optimal classification algorithm. As shown in Table 3, when using the full feature set, the Decision Tree (DT) algorithm has the highest accuracy (97.00%), followed by the Random Forest (RF) algorithm (96.82%). However, the RF has a slightly higher ROC-AUC value (0.998) than the DT algorithm (0.975). This indicates that, although the DT is the best-performing predictive algorithm, the RF algorithm is able to distinguish between classes more effectively. The ROC-AUC is a plot of TP rate against FP rate. The least-performing algorithm is the KNN, with an accuracy of 92.00%.

Table 3. Model performance before feature selection

Classifier	Accuracy	F-measure	ROC-AUC
RF	96.82	0.968	0.998
DT	97.00	0.970	0.975
SVM	93.00	0.930	0.944
KNN	92.00	0.920	0.939

The feature engineering method assesses the relevance of stressor elements and their influence on developing the classification model for predictions. The Chi-square method evaluates the relationship between the dataset's features and the target class variable. As indicated in Table 4, the top four features influencing the predictive model are depression, future career concerns, blood pressure, and self-esteem. The three least influential factors are mental health history, living conditions, and study load.

Table 4. Chi-square Feature Selection - Top 8

Feature	Chi-square value	Rank
depression	1630.64	1
future_career_concerns	1617.97	2
blood_pressure	1611.03	3
self_esteem	1605.76	4
bullying	1559.65	5
sleep_quality	1533.07	6
anxiety_level	1528.25	7
headache	1285.3	8

The model's performance is evaluated post-feature engineering by eliminating the seven ineffective features from the training dataset. As shown in Table 5, following feature engineering, the performance of all classifiers, with the exception of SVM, exhibited a marginal improvement. The DT still emerged as the best-performing algorithm with an accuracy of 97.64%, followed by the RF with an accuracy of 97.00%. The ROC-AUC of RF still outperforms that of DT. After feature selection, the SVM now became the worst-performing algorithm.

Table 5. Model performance after feature selection

Classifier	Accuracy	F-measure	ROC-AUC
RF	97.00	0.970	0.999
DT	97.64	0.976	0.980
SVM	92.91	0.929	0.943
KNN	94.45	0.945	0.957

In finalizing the best performing classifier between RF and DT algorithms, we used the confusion matrix to assess the errors related to cluster predictions after feature engineering. The confusion matrix in Fig. 9 shows that the DT algorithm performs well with Clusters 0 and 1, while the RF algorithm performs better with Cluster 2. In Cluster 0, the DT incorrectly classifies 10 members as belonging to Cluster 2, whilst the RF misclassifies 14 members as Cluster 2. In Cluster 1, the DT erroneously classifies 3 members as belonging to Cluster 2, whereas the RF misclassifies 10 members. In Cluster 2, the DT inaccurately classifies 10 and 3 members as belonging to Clusters 0 and 1, respectively, whereas the RF misclassifies only 9 members as Cluster 0.

The Decision Tree (DT) algorithm from the analysis emerged as the best-performing classifier, followed by the Random Forest (RF) algorithm.

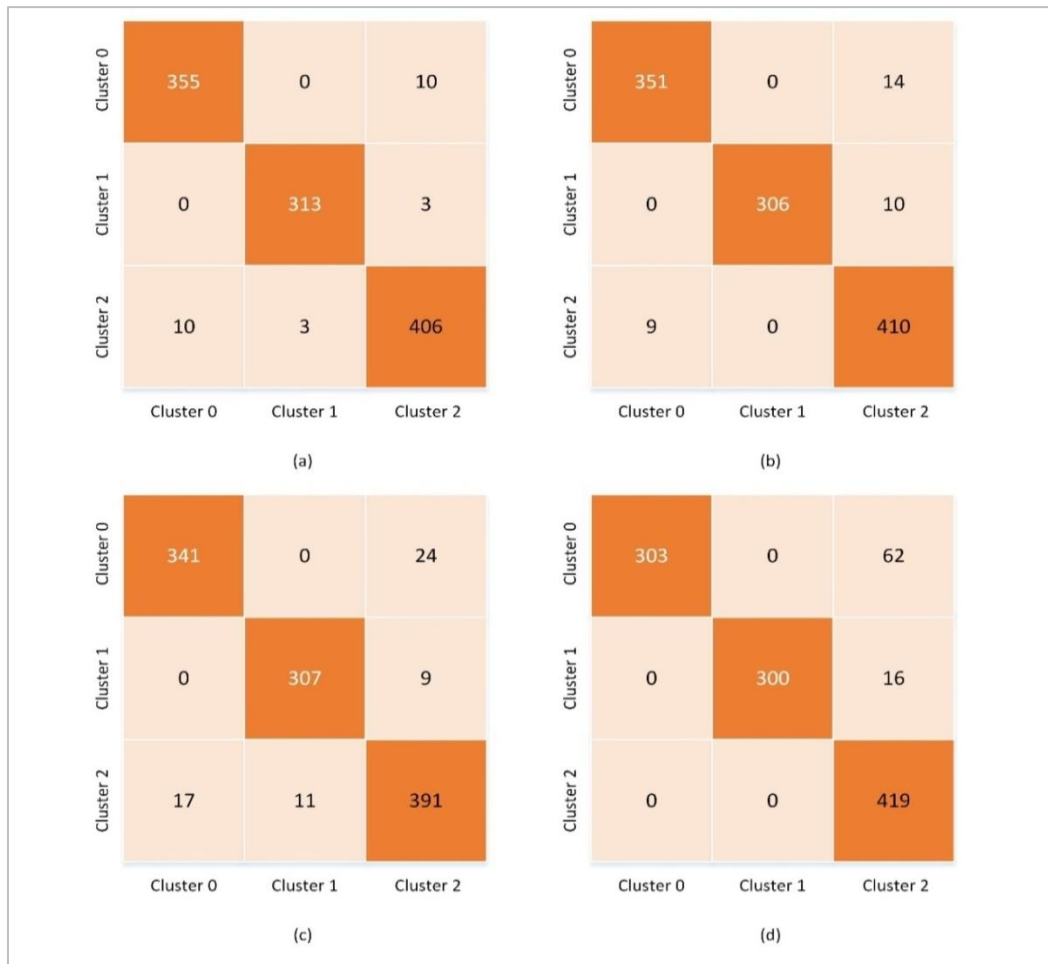


Fig. 9. Confusion Matrix of Best Performing Classifiers after Feature Selection. (a) DT. (b) RF. (c) KNN. (d) SVM

5. Discussion and Findings

The study focused on two key objectives. The first step is to group students based on stressors, and the second is to build a predictive model once the clusters have been formed. Based on the publicly available dataset, we used both the elbow and Silhouette methods to ascertain the number of clusters that will be formed. The PCA and K-means algorithms were used to generate three clusters. The interpretation of the clusters reveals peculiarity among cluster members. The level of risk, or urgency, associated with delivering counselling services varies by cluster. One cluster group required immediate counselling due to the stress factors affecting their lives on campus. The counselling department can frequently meet with the at-risk group or offer an online support system to address their concerns. The at-risk cluster group was affected by almost all the stressors implemented. The counselling department needs a thorough solution that includes the families of this group of students. A social support system or scholarship programmes are fundamental to the solution domain. Any delay in addressing the at-risk cluster group will result in learner retention challenges, poor academic performance, health difficulties, and other undesirable student behaviours. The clustering also reveals another group, which, according to the stressor analysis, comes from relatively stable families. These students live in high-quality campus hostels and receive strong support from their families, and teachers. They maintain a reserved demeanor and lower peer pressure. This group of students needs minimal therapy but should be encouraged to participate in extracurricular activities and socialise more with their peers. Their socialisation can impact at-risk students positively and foster a favourable environment for academic pursuits.

A classification model based on the three clusters was built using four supervised machine learning methods. The classification is crucial in predicting the cluster groupings of future students. This will automate counselling in institutions, eliminating delays caused by a rising student population. Feature engineering was conducted using the Chi-square technique during classification to identify the most significant features for cluster predictions. The primary factors that affected the predictive model significantly are depression, future career concerns, blood pressure, and

self-esteem. The three least impactful features, include mental health history, living conditions, and study load. The feature evaluation resulted in higher accuracy across three of the classification algorithms. The accuracy, F-measure, ROC-AUC, and confusion matrix analyses all show that the DT method has the maximum accuracy for the dataset. This was followed closely by the RF algorithm. However, the SVM algorithm was the least efficient. Research in literature, including works by Firoz et al. [41] and Gorabal et al. [43] for the study and classification of stressors, consistently identifies the DT algorithm as the most effective classifier. The good performance of RF was also documented in literature by Disha Sharma et al. [37], Paul & Resija [39], Rois et al. [5], and Filippis et al. [1]. Even though Filippis et al. [1] used the same dataset for stressor factor engineering, their study focused on feature importance analysis without recourse to clustering or building a predictive model.

6. Conclusion, Limitations and Future Work

Stress among students is evident in several studies in the literature, as summarised in the review section of this paper. Both basic and machine learning studies have pointed to the necessity to analyse stress factors influencing students and provide effective ways for reducing them. This study used K-means and supervised learning algorithms to explore stress-related characteristics in students. K-means clustering was employed to identify stressor groups and create a machine learning system for proactive, rapid, and intelligent group counselling. The results of the unsupervised portion of the study led to the conclusion that cluster analysis for student counselling in an era of Education 4.0 and a continuously increasing student population is essential for quality education and learner retention. Furthermore, the number of clusters formed is determined by the size of the dataset, stressor factor similarities, data type, and the various cluster numbers obtained using the elbow and silhouette methods. The study also examined PCA in K-means, and the visual outcomes indicate that PCA is crucial when the number of factors in cluster formation is substantial. PCA in K-means for cluster analysis brings about cohesion among the cluster groups and a clear separation of one cluster from another. The classification model applied to the dataset, along with the Chi-square method for feature engineering, elucidates the machine learning metrics and performance detailed in our study. The four classifiers, namely DT, RF, SVM, and KNN, demonstrated machine learning performance outcomes corroborated by existing research. The DT emerged as the best-performing algorithm with an accuracy of 97.64% after feature engineering. In the research, it was evident that feature engineering can improve the accuracy of a classifier significantly. In summary, one aspect of education that should be central to educational stakeholders is effective counselling. With the emergence of machine learning, educational institutions ought to prioritise intelligent applications that assist students with personalised counselling services and automate counselling procedures to accommodate the growing student population and mitigate potential delays in addressing learner issues.

Although the study reveals interesting patterns for cluster analysis and prediction, there were some minor setbacks. The dataset size is limited and cannot be used to generalise the stressors impacting students across all higher institutions. A more extensive dataset will enhance the conclusions of this study. Institutional datasets will be distinctive and unique for counselling at different educational institutions. Despite the implementation of four supervised learning algorithms, alternative classifiers may yield superior classification accuracy. The study was unable to examine all supervised learning techniques to reach a comprehensive conclusion based on the dataset. Finally, despite the thoroughness of the literature analysis, certain review articles may not have been included in this study.

In future research, we will compare supervised learning, statistical analysis, and deep learning methods with cluster formation and predictive modelling to understand stressors that mostly affect students.

Author Contributions Statement

The author was responsible for the conceptualization, methodology, data curation, software implementation, formal analysis, model development, validation, visualization, and writing (original draft, review, and editing) of this study.

The author has read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The author declares no conflict of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

This study analysed publicly available datasets. The dataset used in this study is available at <https://github.com/Foysal440/Stress-analysis/blob/main/StressLevelDataset.csv>, accessed on 20th October, 2024

Ethical Declarations

This study used publicly available data and did not involve direct interaction with human participants or animals. Therefore, ethical approval was not required.

Acknowledgments

The author sincerely thanks the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the study and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this work, the author used ChatGPT (OpenAI) to improve the clarity, language, and readability of the manuscript. All generated content was carefully reviewed, edited, and validated by the author. The author takes full responsibility for the content of the publication.

Abbreviations

The following abbreviations are used in this manuscript:

AI – Artificial Intelligence
 ML – Machine Learning
 RF – Random Forest
 DT – Decision Tree
 SVM – Support Vector Machine
 KNN – K-Nearest Neighbor
 PCA – Principal Component Analysis
 ROC-AUC – Receiver Operating Characteristic – Area Under the Curve
 WCSS – Within-Cluster Sum of Squares
 PSS – Perceived Stress Scale
 NB – Naïve Bayes
 MP – Multilayer Perceptron
 GBM – Gradient Boosting Machine
 CGPA – Cumulative Grade Point Average
 SSA – Sub-Saharan Africa

References

- [1] R. De Filippis, · Abdullah, and A. Foysal, “Comprehensive analysis of stress factors affecting students: a machine learning approach,” *Discov. Artif. Intell.* 2024 41, vol. 4, no. 1, pp. 1–17, 2024, doi: 10.1007/s44163-024-00169-6.
- [2] Z. J. Gazzaz *et al.*, “Perceived stress, reasons for and sources of stress among medical students at Rabigh Medical College, King Abdulaziz University, Jeddah, Saudi Arabia,” *BMC Med. Educ.*, vol. 18, no. 1, pp. 1–9, 2018, doi: 10.1186/s12909-018-1133-2.
- [3] World Health Organisation (WHO) (2023) : <https://www.who.int/news-room/questions-and-answers/item/stress>
- [4] United Nations Children's Fund (UNICEF) (2022): <https://www.unicef.org/parenting/mental-health/what-is-stress>
- [5] R. Rois, M. Ray, A. Rahman, and S. K. Roy, “Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms,” *J. Heal. Popul. Nutr.*, vol. 40, no. 1, pp. 1–12, 2021, doi: 10.1186/s41043-021-00276-5.
- [6] Y. Amanvermez *et al.*, “Stress Management Interventions for College Students: A Systematic Review and Meta-Analysis,” *Clin. Psychol. Sci. Pract.*, vol. 30, no. 4, pp. 423–444, 2023, doi: 10.1111/cpsp.12342.
- [7] A. Drigas and E. Mitsea, “Metacognition, Stress – Relaxation Balance & Related Hormones,” *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 9, no. 1, p. 4, 2021, doi: 10.3991/ijes.v9i1.19623.
- [8] Transforming Education (2024): <https://transformingeducation.org/student-stress-statistics/>
- [9] Research (2024): <https://research.com/education/student-stress-statistics>
- [10] D. Owusu, F. Arthur, R. Okyere-Dankwa, O. Affreh, F. Kwame Kumedzro, and R. S. Maison, “Academic stress and burnout among distance education students in a Ghanaian higher education institution,” *Cogent Educ.*, vol. 11, no. 1, p., 2024, doi: 10.1080/2331186X.2024.2334686.
- [11] Y. Amanvermez *et al.*, “Sources of stress among domestic and international students: a cross-sectional study of university students in Amsterdam, The Netherlands,” *Anxiety, Stress Coping*, vol. 37, no. 4, pp. 428–445, 2024, doi: 10.1080/10615806.2023.2280701.
- [12] J. Choi, “Impact of Stress Levels on Eating Behaviors among,” *Nutrients*, vol. 12, no. 5, pp. 1–8, 2020.

- [13] M. de Witte, A. da S. Pinho, G. J. Stams, X. Moonen, A. E. R. Bos, and S. van Hooren, "Music therapy for stress reduction: a systematic review and meta-analysis," *Health Psychol. Rev.*, vol. 16, no. 1, pp. 134–159, 2022, doi: 10.1080/17437199.2020.1846580.
- [14] S. H. Hamaideh, H. Al-Modallal, M. Tanash, and A. Hamdan-Mansour³, "Depression, anxiety and stress among undergraduate students during COVID-19 outbreak and 'home-quarantine,'" *Nurs. Open*, vol. 9, no. 2, pp. 1423–1431, 2022, doi: 10.1002/nop2.918.
- [15] World Bank (2024): <https://www.worldbank.org/en/topic/tertiaryeducation>
- [16] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [17] S. Park *et al.*, "Improving unsupervised image clustering with robust learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. 2, pp. 12273–12282, 2021, doi: 10.1109/CVPR46437.2021.01210.
- [18] S. Križanić, "Educational data mining using cluster analysis and decision tree technique: A case study," *Int. J. Eng. Bus. Manag.*, vol. 12, pp. 1–9, 2020, doi: 10.1177/1847979020908675.
- [19] A. Dutt, "Clustering Algorithms Applied in Educational Data Mining," *Int. J. Inf. Electron. Eng.*, no. March, 2015, doi: 10.7763/ijee.2015.v5.513.
- [20] M. C. S. Áz-Manzanares, J. J. Rodríguez-Déz, J. F. Déz-Pastor, S. Rodríguez-Arribas, R. Marticorena-Sánchez, and Y. P. Ji, "Monitoring of student learning in learning management systems: An application of educational data mining techniques," *Appl. Sci.*, vol. 11, no. 6, 2021, doi: 10.3390/app11062677.
- [21] S. Tiwari, "Supervised Machine Learning: A Brief Introduction," *Proc. Int. Conf. Virtual Learn.*, vol. 17, no. 5, pp. 219–230, 2022, doi: 10.58503/icvl-v17y202218.
- [22] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.
- [23] A. Khalaf Hamoud *et al.*, "Supervised Learning Algorithms in Educational Data Mining: A Systematic Review," *Southeast Eur. J. Soft Comput.*, vol. 10, no. 1, pp. 55–70, 2021, [Online]. Available: <http://scjournal.ius.edu.ba/index.php/scjournal/article/view/199>.
- [24] B. Albreiki, N. Zaki, and H. Alashwal, "Revisión bibliográfica sistemática de la predicción del rendimiento de los estudiantes mediante técnicas de aprendizaje automático," *Educ. Sci.*, vol. 11, no. 9, 2021.
- [25] Y. Deng *et al.*, "Family and Academic Stress and Their Impact on Students' Depression Level and Academic Performance," *Front. Psychiatry*, vol. 13, no. June, pp. 1–13, 2022, doi: 10.3389/fpsy.2022.869337.
- [26] A. Alkhalwaldeh *et al.*, "Stress Factors, Stress Levels, and Coping Mechanisms among University Students," *Sci. World J.*, vol. 2023, 2023, doi: 10.1155/2023/2026971.
- [27] W. Thanoi, N. Vongsirimas, Y. Sitthimongkol, and P. Klainin-Yobas, "Examining Predictors of Psychological Well-Being among University Students: A Descriptive Comparative Study across Thailand and Singapore," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, 2023, doi: 10.3390/ijerph20031875.
- [28] S. Asif, A. Mudassar, T. Z. Shahzad, M. Raouf, and T. Pervaiz, "Frequency of depression, anxiety and stress among university students," *Pakistan J. Med. Sci.*, vol. 36, no. 5, pp. 971–976, 2020, doi: 10.12669/pjms.36.5.1873.
- [29] K. J. Reddy, K. R. Menon, and A. Thattil, "Academic stress and its sources among university students," *Biomed. Pharmacol. J.*, vol. 11, no. 1, pp. 531–537, 2018, doi: 10.13005/bpj/1404.
- [30] D. I. Khelil, D. Z. S. A. Altarawneh, and D. M. A. S. Khasawneh, "Thinking Patterns and their Impact on Coping with Psychological Stress among Jordanian Public University Students," *Migr. Lett.*, vol. 20, no. 6, pp. 1089–1102, 2023, doi: 10.59670/ml.v20i6.4920.
- [31] F. G. Karaoglan Yilmaz, R. Yilmaz, and F. Erdogan, "The Relationship Between Nomophobia, Emotional Intelligence, Interpersonal Problem-Solving, Perceived Stress, and Self-esteem Among Undergraduate Students," *J. Technol. Behav. Sci.*, vol. 8, no. 1, pp. 27–35, 2023, doi: 10.1007/s41347-022-00282-2.
- [32] M. R. Shehada, G. Alfakhry, I. Jamous, A. A. Aljoujou, and M. Abdulhak, "Major Stress Sources Amongst Dental Students at Damascus University, Syria," *Int. Dent. J.*, vol. 73, no. 2, pp. 205–211, 2023, doi: 10.1016/j.identj.2022.03.005.
- [33] C. E. Brett, M. L. Mathieson, and A. M. Rowley, "Determinants of wellbeing in university students: The role of residential status, stress, loneliness, resilience, and sense of coherence," *Curr. Psychol.*, vol. 42, no. 23, pp. 19699–19708, 2023, doi: 10.1007/s12144-022-03125-8.
- [34] L.-A. Hanna, M. Wilson, M. Hall, and A. Hanna, "A Questionnaire Study to Investigate Stress among Future Pharmacists by Gender and Year Group," *Pharmacy*, vol. 6, no. 3, p. 75, 2018, doi: 10.3390/pharmacy6030075.
- [35] J. M. Ngaboyishema, S. Gisanabagabo, J. Paul, and K. Kayonde, "Perception of Stress among Masters' Students at Adventist University of Central Africa (AUCA)," vol. 6, pp. 83–92, 2022.
- [36] A. Pitt, F. Opreacu, G. Tapia, and M. Gray, "An exploratory study of students' weekly stress levels and sources of stress during the semester," *Act. Learn. High. Educ.*, vol. 19, no. 1, pp. 61–75, 2018, doi: 10.1177/1469787417731194.
- [37] D. S. *et al.*, Disha Sharma *et al.*, "Stress Prediction of Students using Machine Learning," *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 5609–5620, 2020, doi: 10.24247/ijmperdjun2020534.
- [38] K. Bhatia, M. Kumar, R. K. Bhatia, A. Sharma, and K. Chhabra, "Machine learning based classification of academic stress factors," *Proc. - 2021 4th Int. Conf. Comput. Intell. Commun. Technol. CCICT 2021*, pp. 45–50, 2021, doi: 10.1109/CCICT53244.2021.00020.
- [39] M. A. Paul and M. Resija, "Stress Detection in College Students Using Machine Learning Algorithm," *Int. J. Res. Publ. Rev. J. homepage www.ijrpr.com*, vol. 5, no. 2, pp. 3411–3417, 2024, [Online]. Available: www.ijrpr.com.
- [40] R. Ahuja and A. Banga, "Mental stress detection in university students using machine learning algorithms," *Procedia Comput. Sci.*, vol. 152, pp. 349–353, 2019, doi: 10.1016/j.procs.2019.05.007.
- [41] M. Firoz, M. M. Islam, M. Shidujaman, A. Islam, and M. T. Habib, "University student's mental stress detection using machine learning," vol. 12779, no. Icmir 2023, p. 113, 2023, doi: 10.1117/12.2690039.
- [42] C. Ding, Y. Zhang, and T. Ding, "A systematic hybrid machine learning approach for stress prediction," *PeerJ Comput. Sci.*, vol. 9, pp. 1–25, 2023, doi: 10.7717/PEERJ-CS.1154.

- [43] Dr. J V Gorabal, Mr. Vipul Y S, Mr. Yashas S Bharadwaj, Mr. Mohammed Anas, and Mr. Mohammed Nehad Ameer, "Predictive Analysis of Student Stress Level using Machine Learning," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 12, no. 05, pp. 465–473, 2023, doi: 10.48175/ijarsct-10387.
- [44] R. Madhuri, M. R. Murty, J. V. R. Murthy, P. V. G. D. P. Reddy, and S. C. Satapathy, "Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms," *Adv. Intell. Syst. Comput.*, vol. 249 VOLUME, pp. 137–144, 2014, doi: 10.1007/978-3-319-03095-1_15.
- [45] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electron.*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.
- [46] Global Tech Council (2020): <https://www.globaltechcouncil.org/data-science/k-means-clustering-vs-hierarchical-clustering/>
- [47] Blockchain Council (2024): <https://www.blockchain-council.org/blogs/difference-between-k-means-and-hierarchical-clustering/>
- [48] J. Xu and K. Lange, "Power k-means clustering," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 11977–11991, 2019.
- [49] T. M. Ghazal *et al.*, "Performances of k-means clustering algorithm with different distance metrics," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.
- [50] M. Cui, "on the Elbow Method," pp. 5–8, 2020, doi: 10.23977/accaf.2020.010102.
- [51] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review," *Front. Aging Neurosci.*, vol. 9, no. OCT, pp. 1–12, 2017, doi: 10.3389/fnagi.2017.00329.
- [52] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, 2021, doi: 10.1088/1742-6596/1817/1/012009.
- [53] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. November 2021, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.
- [54] Y. Zhong, "The analysis of cases based on decision tree," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 0, pp. 142–147, 2016, doi: 10.1109/ICSESS.2016.7883035.
- [55] C. Hu, C. Zhang, D. Lei, T. Wu, X. Liu, and L. Zhu, "Achieving Privacy-Preserving and Verifiable Support Vector Machine Training in the Cloud," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 3476–3491, 2023, doi: 10.1109/TIFS.2023.3283104.
- [56] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-2941-7.
- [57] P. Thanh Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery," *Sensors (Basel)*, vol. 18, no. 1, 2017, doi: 10.3390/s18010018.
- [58] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.
- [59] M. Mehedi Hassan, S. Mollick, and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection," *Healthc. Anal.*, vol. 2, no. September, p. 100112, 2022, doi: 10.1016/j.health.2022.100112.
- [60] C. L. Clayman, S. M. Srinivasan, and R. S. Sangwan, "K-means clustering and principal components analysis of microarray data of L1000 landmark genes," *Procedia Comput. Sci.*, vol. 168, no. 2019, pp. 97–104, 2020, doi: 10.1016/j.procs.2020.02.265.
- [61] M. K. Islam, M. S. Ali, M. S. Miah, M. M. Rahman, M. S. Alam, and M. A. Hossain, "Brain tumor detection in MR image using superpixels, principal component analysis and template based K-means clustering algorithm," *Mach. Learn. with Appl.*, vol. 5, no. January, p. 100044, 2021, doi: 10.1016/j.mlwa.2021.100044.

Author's Profile



Delali Kwasi Dake is an Associate Professor of Computing and Information Technology at the University of Education, Winneba, Ghana and currently the Head, Department of ICT Education. He holds a PhD in Computer Engineering from the Kwame Nkrumah University of Science & Technology, Kumasi, Ghana. His research focuses on educational data mining, artificial intelligence, reinforcement learning, intelligent systems, software-defined networks, and blockchain technology.

Prof. Dake is currently an external examiner for KNUST, BlueCrest, and Heritage Christian University. He is an academic reviewer for highly indexed journals, including Signal, Image and Video Processing – Springer; Scientific Reports - Springer Nature; Journal of Network and Systems - Springer; Education and Information Technologies - Springer; Cluster Computing - Springer; Discover Computing - Springer; BMC Medical Informatics and Decision Making - Springer; Frontiers of Digital Education - Springer; BMC Oral Health, Data Science Aspect - Springer; MPDI Education Sciences - Scopus indexed; and PeerJ Computer Science - Scopus indexed.

How to cite this paper: Delali Kwasi Dake, "Cluster Pattern Analysis of Students Stress using Machine Learning Algorithms with Feature Engineering", *International Journal of Modern Education and Computer Science(IJMECS)*, Vol.18, No.2, pp. 147-162, 2026. DOI:10.5815/ijmeecs.2026.02.09