Forecasting the Psychological Well-being of Engineering Students Utilizing Machine Learning

1st Mr.Sangram Lembe
Department of Data Science and Business System
S.R.M. Institute of Science And Technology
Chennai, India.
sangramlembe@gmail.com

2nd Mr. Sanchit Chemate
Master in Human Resource Development
Jamnalal Bajaj Institute of Management Studies
Mumbai, India
sanchitchemate@gmail.com

***Abstract*—The rising global prevalence of mental health disorders, coupled with the rigorous demands of engineering curricula, underscores an urgent need for proactive mental health support among engineering students. The pressures associated with intensive coursework, high-stakes assessments, and competitive job markets profoundly affect the psychological health of this demographic. This project aimed to forecast the psychological well-being of engineering students by identifying and analyzing the most impactful influencing factors using machine learning. A comprehensive psychological assessment was conducted via a structured questionnaire, meticulously designed in consultation with specialist psychologists. The collected data underwent extensive preprocessing, including the standardization of 67 numerical columns using Z-score normalization. For predictive modeling, a stacked ensemble learning approach was implemented, utilizing five base models (Logistic Regression, K-Nearest Neighbours, Classification and Regression Trees, Support Vector Machines, and Naive Bayes). After rigorous performance comparison, the Multi-Layer Perceptron (MLP) was selected as the optimal meta-learner. Optimization and feature selection were enhanced using techniques such as ANOVA, Mutual Information, Recursive Feature Elimination, Principal Component Analysis, and Particle Swarm Optimization (PSO). The results demonstrated the efficacy of the stacked ensemble model, with the MLP meta-model achieving the highest stacking accuracy of 90.8%. Among individual classifiers, the MLP model achieved the highest overall performance with an F1 Score of 96.52%. Furthermore, feature selection through PSO yielded an accuracy of 96.33%. Key factors consistently identified as highly influential in predicting well-being included stress management, routine pattern, and engineering passion. This research highlights the practical advantage of advanced data analytics in proactively identifying and supporting stressed students, enabling targeted interventions and mitigating adverse effects on academic performance and holistic development.**

***Index Terms*—Engineering, Students, Psychological Well-being, Machine Learning**

I. Introduction

The rising global prevalence of mental health disorders, coupled with the rigorous demands of engineering curricula, underscores an urgent need for proactive mental health support among engineering students. The pressures associated with intensive coursework, high-stakes assessments, and competitive job markets profoundly affect the psychological health of this demographic. This project aimed to forecast the psychological well-being of engineering students by identifying and analyzing the most impactful influencing factors using machine learning. A comprehensive psychological assessment was conducted via a structured questionnaire, meticulously designed in consultation with specialist psychologists. The collected data underwent extensive preprocessing, including the standardization of 67 numerical columns using Z-score normalization. For predictive modeling, a stacked ensemble learning approach was implemented, utilizing five base models (Logistic Regression, K-Nearest Neighbours, Classification and Regression Trees, Support Vector Machines, and Naive Bayes). After rigorous performance comparison, the Multi-Layer Perceptron (MLP) was selected as the optimal meta-learner. Optimization and feature selection were enhanced using techniques such as ANOVA, Mutual Information, Recursive Feature Elimination, Principal Component Analysis, and Particle Swarm Optimization (PSO). The results demonstrated the efficacy of the stacked ensemble model, with the MLP meta-model achieving the highest stacking accuracy of 90.8%. Among individual classifiers, the MLP model achieved the highest overall performance with an F1 Score of 96.52%. Furthermore, feature selection through PSO yielded an accuracy of 96.33%. Key factors consistently identified as highly influential in predicting well-being included stress management, routine pattern, and engineering passion. This research highlights the practical advantage of advanced data analytics in proactively identifying and supporting stressed students, enabling targeted interventions and mitigating adverse effects on academic performance and holistic development.

II. Literature Review

Psychological assistance among engineering graduates is evident due to academic, professional, and personal challenges they face. Early detection and tailored interventions are crucial for enhancing their mental health and overall success. A range of studies have explored the use of machine learning in predicting mental health issues and academic performance among students. Mutalib, S., Rahim and Shafiee, N. S. M.,

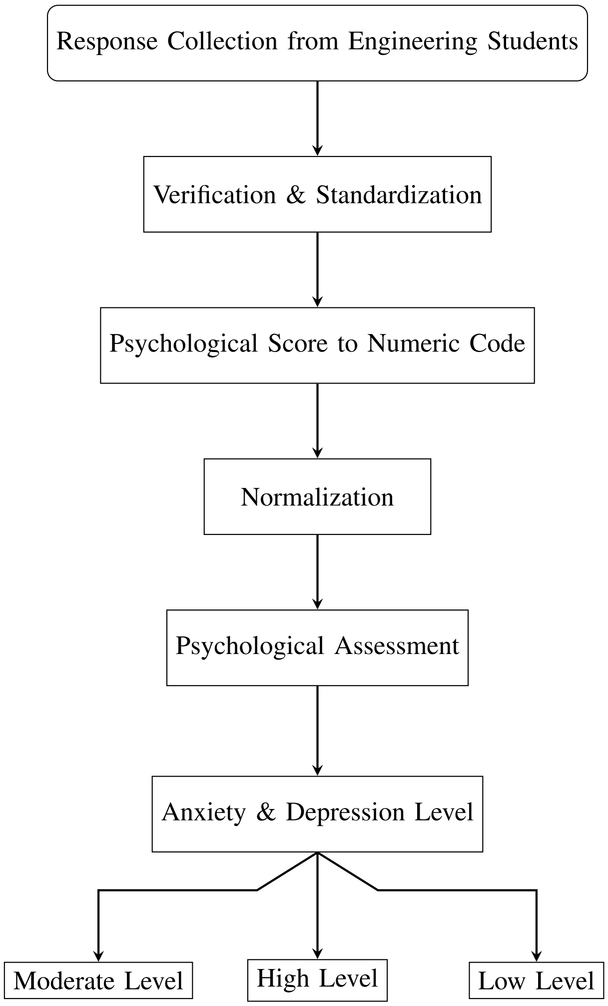

Fig. 1: Proposed Methodology for Assessing Student Psychological Levels.

Zamani, M., Mohammad both focused on identifying factors and classifying mental health problems, with Mutalib finding Decision Tree to be the most accurate model for stress, depression, and anxiety.Gorabal, J. V., Jonnadula emphasized the need for a comprehensive system to predict and manage student stress, Halde, R. R., Deshpande demonstrated the impact of student psychology on academic performance, with machine learning algorithms improving prediction accuracy. These studies collectively highlight the potential of machine learning in profiling students for career prediction, by considering their mental health and academic performance.

The study by S. Mutalib (2021) utilizes machine learning models to predict mental health issues among higher education students. Decision Tree, Support Vector Machine (SVM), and Neural Network show the highest accuracy for stress, depression, and anxiety, respectively. The methodology involves collecting survey data, applying various machine learning algorithms, and evaluating model performance based on accuracy, sensitivity, specificity, and precision. Decision Tree and SVM perform best for stress and depression, while Logistic Regression and Neural Networks yield fair results for anxiety. [7]

Dr. J V Gorabal's study (2023) focuses on predicting student stress levels using machine learning techniques. While it accurately predicts stress levels, it lacks in uncovering underlying relationships within psychological data. The study proposes implementing a comprehensive system to manage student stress by providing personalized suggestions. [5]

Radhika R. Halde's study (2016) explores psychology-assisted prediction of academic performance using machine learning. While it demonstrates a 4 to 6 percent increase in prediction accuracy, limitations include focusing solely on final year students, reliance on self-reported data for variables like motivation, and failure to consider external factors influencing academic performance. The study employs Neural Network for numeric prediction of CGPA and Decision Tree for classification of failures in the sixth semester. [4]

Nor Safika Mohd Shafiee's study (2020) examines mental health issues among higher education students and reviews existing machine learning techniques for prediction. It discusses contributing factors to these problems but does not address specific challenges in their identification [12].

Jetli Chung's study (2022) discusses machine learning approaches for predicting mental health problems and their application in psychological profiling of students for career prediction. While it provides a systematic literature review and identifies challenges, limitations, and potential opportunities, it has limitations including small sample sizes and insufficient validation from external sources. The methodology involved searching reliable databases, adhering to the PRISMA methodology, and categorizing identified articles. Support Vector Machine, Random Forest, and Artificial Neural Networks were utilized [2].

Marouf's study (2019) uses machine learning to predict psychological behavior and stress reactions in undergraduate students with around 70 percent accuracy. It highlights limitations in sample specificity and potential bias from self-reported data. The Sequential Minimal Optimization (SMO) and k-NN classifier models achieve the highest accuracy. Data collection involved 150 participants, assessing personality traits and stress levels, and employing nine machine learning models for evaluation [6].

Ravinder Ahuja's study (2019) focuses on detecting mental stress in university students using machine learning algorithms. Despite limitations such as data limited to one institute and a small sample size of 206 students, Support Vector Machine achieves the highest accuracy. The study analyzes stress before exams and during internet usage, correlating it with online time. It highlights the importance of addressing factors like exam pressure and recruitment stress. Methodology involves applying four classification algorithms and enhancing accuracy through 10-Fold Cross-Validation. Identified research gaps include the impact of exam pressure and recruitment stress on mental stress levels and the correlation between stress factors and internet usage time [1].

A study by Rahman et al. (2022) employed machine learning algorithms, including random forest and adaptive boosting, to accurately identify negative mental well-being traits among college students using health behavior data. Key findings include the identification of important predictors such as body mass index, sports activities per week, GPA, sedentary hours, and age. The study highlighted the superiority of random forest over other classifiers like k-nearest neighbor and naïve-Bayes. Limitations include the cross-sectional design, self-reporting bias, and unavailable data due to intellectual property regulations. The study underscores the need for more sophisticated methods, such as natural language processing, and emphasizes the scarcity of similar studies in Asian populations. [10]

Chiyue Wang's study presented at the 2023 AIKIIE conference delves into mental health situations among college students using machine learning algorithms. The research underscores the necessity for tailored intervention strategies based on gender, academic year, and personal background. While successful in analyzing mental health data and identifying critical factors, the study acknowledges limitations in generalizability and potential biases introduced by machine learning algorithms. Despite not detailing specific interventions, the study aims to inform the development of targeted mental health programs. Methodologically, the research employs various machine learning algorithms and advanced techniques to dissect large-scale data, ultimately contributing insights for crafting effective interventions in educational settings. [16]

Yutao Sun's study presented at the 2021 IEEE conference addresses the potential of machine learning for analyzing and warning about college students' psychological data, aiming to enhance mental health education. Identified limitations include the absence of comparative studies on machine learning algorithms, lack of established algorithm evaluation metrics for practical problems, and the need for further exploration into the application of training models in real-world scenarios. The study collected real datasets from a Chinese university and tested ten widely used machine learning algorithms to determine the optimal model for mental health screening. Methodologically, the research involved testing various algorithms on collected datasets and designing a prototype system for early warning of mental health issues. Despite limitations in real-world implementation and potential contextual restrictions. [15]

In the 2023 paper by Akash D published in the International Journal of Scientific Research in Engineering and Management, a system is proposed for predicting stress levels among college students using machine learning (ML) algorithms. This system not only identifies stress but also provides solutions for alleviating it. Akash D identifies shortcomings in existing stress management methods and institutional approaches to tracking and mitigating student stress, stressing the need for improved solutions. The study employs ML algorithms like Naive Bayes, K-Nearest Neighbor, and Decision Tree, along with the SVM technique and Bayes classifier, for stress assessment and solution provision. Methodologically, data is collected through survey forms, and techniques like SVM and Bayes classifier are used for data classification and analysis. The findings highlight the potential benefits of ML-based stress prediction and solutions in reducing student stress levels. However, the study suggests further research to analyze the relative efficiency of different algorithms and conduct a comparative study. [3]

In the 2023 CyMaEn conference paper by Pornrat Nison et al., a machine-learning approach for detecting depression in college students is presented. It utilizes general information without direct mental health questions and addresses class imbalance with resampling methods. The approach achieves an average predicting accuracy of up to 0.66, indicating progress in machine learning-based depression screening for college students. [8]

In a 2023 conference paper by Yang Zhen, a model for predicting college students' mental health based on status data is introduced. The study employs the Apriori algorithm and XGBoost model, enhanced with the SMOTE+ENN method, achieving high precision, recall, F1 score, and AUC. With a precision of 0.87, recall of 0.86, F1 score of 0.86, and AUC of 0.89, the proposed model outperforms traditional machine learning and deep learning approaches. Methodologically, features are extracted from students' status data, and the Apriori algorithm with pre-pruning is applied, followed by the use of the XGBoost model optimized by SMOTE+ENN for prediction. [17]

In the 2022 ASSIC conference paper by Minakshi Roy et al., multimodal machine learning approaches for career prediction are explored. The study utilizes machine learning algorithms including ADABOOST, Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) to predict students' career choices. Results indicate that SVM achieves the highest accuracy of 98 percent , outperforming other algorithms. However, the study lacks discussion on limitations associated with computational approaches and does not specify limitations for the algorithms used. Methodologically, the dataset is trained and tested with the aforementioned algorithms to predict career choices. [11]

In the 2020 paper published in the International Journal of Recent Technology and Engineering, the use of machine learning algorithms for predicting student performance and career paths is discussed. The study highlights limitations including a limited dataset obtained from a single university, which may not represent the entire student population, and the need for future research to collect data from multiple universities for better generalization. Methodologically, various machine learning algorithms such as Linear Regression, Logistics Regression, Support Vector Machine, Naïve Bayes Classifier, and K-means Clustering are applied to predict student career paths, with their efficiencies compared using real data from university students. [13]

In the 2023 Journal of Advanced College of Engineering and Management, Subiddhya Panthee et al. present a Career Guidance System using machine learning algorithms to predict suitable career paths for students based on their skills, interests, and personality traits. The system collects data from webbased questionnaires to predict users' Big Five personality traits and VAK learning types, employing algorithms such as Random Forest, Support Vector Machine, K-Nearest Neighbor, K-Means Clustering, and Artificial Neural Networks. Methodologically, the study utilizes data mining techniques and preprocessing methods to enhance decision-making in career selection, though it faces limitations including reliance on a poor quality dataset for the VAK learning model and overfitting on training data of age. [9]

M. Srividya's study (2018) utilizes machine learning algorithms to predict mental health states in various groups, including students, for career prediction purposes. The study emphasizes the importance of positive mental health for productivity and potential realization. Methodology involves applying machine learning algorithms, including support vector machines, decision trees, and naïve Bayes classifier, to identify mental health states in different target groups. Cluster labels are validated using the Mean Opinion Score, and classifiers for mental health prediction are built. [14]. Based on our literature review, we identified the following issues that must be addressed:

i. Reliance on self-reported data from one institute limits generalizability and may introduce bias.
ii. Limited focus on final year students and specific stressors overlooks broader mental health concerns.
iii. Failure to consider external factors and reliance on self-reported data may affect accuracy.
iv. Absence of intervention assessment and limited discussion on long-term implications.
v. Small sample size and variability in algorithm performance impact reliability.
vi. Challenges in applying machine learning to real-world practices require further research and improvement.
vii. Lack of discussion on computational limitations and unspecified drawbacks of the utilized algorithms.
viii. Limited by the absence of direct mental health questions and potential biases due to class imbalance.

In conclusion, addressing the identified issues is imperative for advancing the understanding and support of mental health among students. Overcoming limitations such as reliance on self-reported data, narrow focus on specific stressors, and small sample sizes will enhance the generalizability and accuracy of research findings. Furthermore, evaluating interventions and considering long-term implications are essential for developing effective strategies. Continued research to improve the application of machine learning in real-world settings is necessary to provide comprehensive support for student mental health.

Below is the revised methodology section for your research paper, focusing on forecasting the psychological well-being of engineering students using an ensemble learning-based machine learning model. The section has been reviewed for sequence, accuracy, coherence, and proper information

III. Methodology

A. Introduction

Mental health is a critical aspect of overall well-being, and its importance cannot be overstated, especially among students pursuing demanding academic programs like engineering. The rigorous nature of engineering curricula, coupled with various personal and social factors, can contribute to heightened stress levels and potential mental health issues. However, many students may not be aware of their mental health condition or lack the necessary resources to assess these concerns effectively.

This project aims to conduct a comprehensive psychological assessment among engineering students, focusing on identifying the most impactful factors influencing their mental health. By analysing these factors, the study seeks to determine whether a student is experiencing stress or other mental health challenges. The ultimate goal is to provide engineering students with valuable insights and a better understanding of their mental well-being.

Recognizing the significance of mental health in academic and personal life, this project endeavours to offer a tailored and accessible approach to psychological assessment, catering specifically to the unique needs and challenges faced by engineering students. By shedding light on this crucial aspect, the study aims to contribute to the awareness and understanding of mental health among engineering students

B. Data collection

TABLE I: Demographic Information of Engineering Students

Age in Years	Gender	Year of Engineering	Location
18-23	Male (62.8%) Female (37.2%)	1st Year (16.8%) 2nd Year (57.2%) 3rd Year (22%) Final Year (4%)	Rural (53.2%) Urban (46.8%)

TABLE II: Academic Engagement & Career Readiness

Risk Factors	Response in Percentages (%)
Academic satisfaction	Very satisfied: 45% Neutral: 30% Not satisfied: 25%
Extracurricular participation	Yes, frequently: 20% Occasionally: 45% No: 35%
Intellectual level	Advanced: 40% Intermediate: 30% Beginner: 30%
Fear of success	Yes: 25% No: 75%
Job flexibility	Choose another field: 35% Stick with engineering: 65%

To gather comprehensive data for this psychological assessment, a structured questionnaire was meticulously designed in consultation with specialist psychologists and peer counsellors. The questionnaire aimed to capture a wide range of factors potentially influencing the mental health of engineering students, including psychological aspects, background information, interests, engagement levels, interactions, and academic-related elements.

The questionnaire encompassed questions pertaining to students' psychology, exploring their thought processes, emotional states, and coping mechanisms. Additionally, it delved into their personal backgrounds, seeking insights into familial, social, and environmental factors that might contribute to their overall well-being.

Furthermore, the questionnaire investigated students' academic performance, pressure, expectations, and engagement levels within their engineering programs. This allowed for a holistic understanding of the interplay between academic demands and mental health.

To ensure a diverse and representative sample, the questionnaire was distributed as a Google Form to students across all academic years and departments within the engineering school. Participation was voluntary, and all questions were mandatory, ensuring that each respondent provided comprehensive and candid responses.

TABLE III: Survey Responses on Various Risk Factors

Risk Factor	**Response in Percentages (%)**
1. Financial Stability	Yes: 20.8% No: 79.2%
2. Loneliness	Same city: 20.1% Different district: 75.6% Different state: 3.4% NIR: 0.9%
3. Stress/Depression	Rarely: 16.4% Sometimes: 50.3% Frequently: 33.3%
4. Overall Feelings	Mostly good: 32.5% Okay: 49.8% Mostly bad: 13.9% Really bad: 3.8%
5. Career Backup Plan	Yes: 69.2% No: 30.8%
6. First in Family to College	Yes: 27.1% No: 72.9%
7. Enjoyment of Coursework	A lot: 40.4% Somewhat: 50.5% Not at all: 9.1%
8. Abandoning Hobbies	Yes: 44.4% No: 55.6%
9. Engineering Isolation	Frequently: 30.4% Occasionally: 69.6%
10. Mental Health Challenges	Yes: 79.4% No: 20.6%
11. Coping with Stress	Exercise: 21.4% Talking: 60.6% Ignoring it: 18%
12. Pride in Accomplishments	Yes: 52.8% So-so: 47.2%
13. Regular Physical Activity	Yes: 59.9% No: 40.1%
14. Pursuing Postgraduate	Yes: 42.2% Undecided: 57.8%

By adopting this data collection approach, the study aimed to capture raw and authentic responses from engineering students, reflecting their genuine experiences, challenges, and perspectives. This naturalistic approach was crucial to obtain an accurate and unbiased understanding of the mental health landscape among the target population.

The collected responses were securely stored in Excel spreadsheets, ensuring data integrity and facilitating further analysis and interpretation.

C. Data Prepossessing

The data collected through the Google Form responses constituted the initial dataset for this study. However, before proceeding with further analysis, several prepossessing steps were undertaken to ensure the data integrity, completeness, and suitability for subsequent modelling and interpretation. Firstly, the dataset contained a mix of categorical and numerical data. To facilitate analysis, categorical data were converted into numerical form using an encoding technique. In consultation with counsellors and psychologists, certain questions and responses deemed irrelevant or repetitive were excluded from the dataset to eliminate redundancy and enhance data quality.

Subsequently, the dataset was thoroughly examined for missing values, and appropriate strategies were employed to address them. If a column or feature had a substantial number of missing values, rendering it insignificant for the analysis, it was removed from the dataset. Columns with a relatively low proportion of missing values were retained for analysis.

Furthermore, to transform the raw data into a meaningful and interpretable form, feature engineering was performed. In close collaboration with counsellors and psychologists, the questions and responses were carefully analysed, and appropriate labels were assigned to each feature, ensuring meaningful and descriptive representations. This process involved grouping related questions and responses into coherent features that could effectively capture the underlying psychological constructs and mental health aspects.

Through these data preprocessing steps, the initial raw dataset was transformed into a structured and refined dataset, ready for further analysis and modelling. The involvement of counsellors and psychologists throughout the process ensured that the data retained its psychological relevance and integrity, facilitating accurate insights into the mental health of engineering students.

D. Normalization

In our dataset comprising 67 numerical columns, normalization played a pivotal role in standardizing the range and distribution of the data for enhanced analysis. Before normalization, we conducted thorough preprocessing to handle missing values and outliers, ensuring the integrity of our data. Employing the Z-score normalization technique, each numerical column was independently standardized to have a mean of 0 and a standard deviation of 1. This process involved subtracting the mean of each column from its data points and dividing by the standard deviation. Post-normalization, we evaluated the distribution and statistical properties of the data through visualizations like histograms and box plots, confirming the effectiveness of the normalization. By mitigating the influence of differing scales and variances across features, normalization facilitated more meaningful comparisons and interpretations of the data, laying a robust foundation for subsequent analyses and insights.

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \text{(Scaling)}$$

E. Model Selection

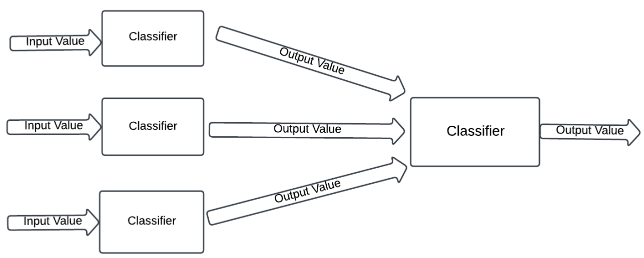

Fig. 2: System Architecture Diagram

After exploring various machine learning models, we decided to use a stacked ensemble for our project. We chose the stacked ensemble because it holds the strengths of each individual model, the ensemble offers a powerful and flexible approach to predicting mental health outcomes, a perfect fit for our research needs.

Understanding Stacked Ensemble Model Imagine you have a group of friends, each with their own expertise. You might ask one friend for advice on sports, another for help with cooking, and so on. Similarly, a stacked ensemble model combines the predictions of different machine learning models, each specialising in its own area, to make a smarter decision.

How Stacked Ensemble Works

Think of the stacked ensemble like a team of models. Each model (or base learner) has their own unique strengths (or machine learning algorithm). When faced with a challenge, the models combine their abilities to deal with issues (or make a prediction). But there's also a leader (or meta learner) who coordinates each and every decision and decides the best course of action based on their input.

Key Advantages of Stacked Ensemble

i. Teamwork: By bringing together different models, the stacked ensemble can solve problems that no single model could tackle alone. Strength in Diversity: The ensemble benefits from the varied approaches of its base learners.
ii. Adaptability: The ensemble can adjust its strategy based on the situation, making it versatile and effective in different scenarios.
iii. Easy to Understand: Despite its complexity, the concept of combining predictions is intuitive and easy to grasp.

F. Model analyzing and training

In the beginning, we created a powerful model by combining different types of algorithms, like Logistic Regression, K-Nearest Neighbours, Decision Trees, and others. This combined model is called a stacked ensemble, which is like having a team of experts working together to solve a problem.

We used a special algorithm, called Support Vector Machine (SVM), to oversee how these different algorithms work together and make the final predictions. Think of SVM as the meta-learner who decides on the best strategy based on everyone's input.

After training our stacked ensemble model, we carefully checked how well it performed. We wanted to make sure it gave accurate predictions, so we tested it thoroughly using different methods.

During testing, we realised we could improve our model even more by trying different strategies for the team captain role. We looked at other algorithms like Decision Trees, Logistic Regression, and Multi-Layer Perceptron (MLP). After comparing their performance, we found that MLP was the best choice. It showed the most promise in making accurate predictions.

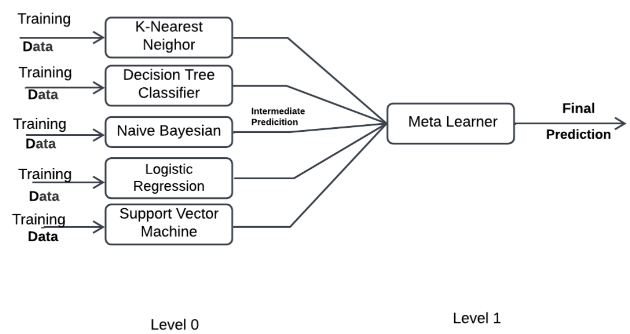

Fig. 3: Model Diagram: Stack Ensemble

So, we decided to use MLP as the new meta-learner for our model. This decision marked an important step in our analysis, indicating that we had finished choosing the best model setup. Now, we're ready to move forward and fine-tune our model to make it even better. .

G. Optimization and Feature Selection Techniques

In the feature selection, we utilised a combination of statistical and optimization techniques to identify and select the most relevant features for our machine learning model.

We began by employing statistical techniques such as ANOVA, MI, and RFE. ANOVA, or Analysis of Variance, is a statistical method used to assess the variance between different groups of data. In the context of feature selection, ANOVA helps identify features that exhibit significant variations across different classes or groups, indicating their potential relevance for predicting the target variable.

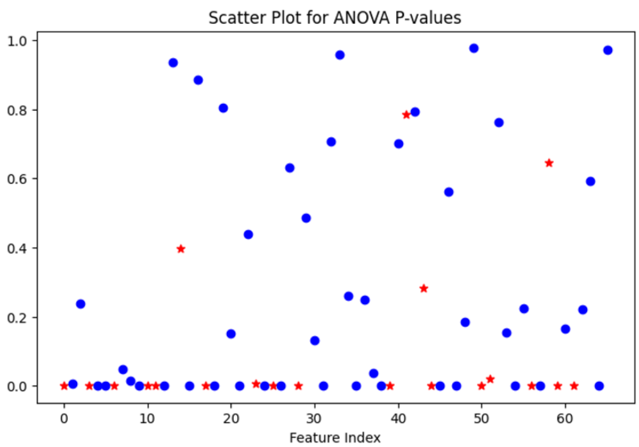

Fig. 4: Scatter Plot ANOVA

Mutual Information (MI) is another statistical technique we employed, measuring the dependency between two variables by quantifying the amount of information obtained about one variable through the other. MI evaluates the information gain of each feature with respect to the target variable, selecting features that provide the most relevant information for prediction.

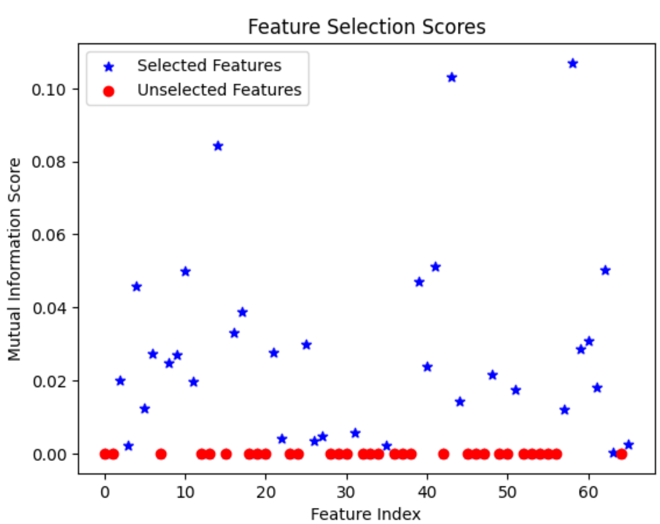

Fig. 5: Feature Selection Score

Recursive Feature Elimination (RFE) is an iterative technique that systematically removes the least important features from the dataset. RFE typically utilises a machine learning model to evaluate feature importance iteratively. It removes features with the least contribution to the model's performance, retaining only the most informative features for prediction.

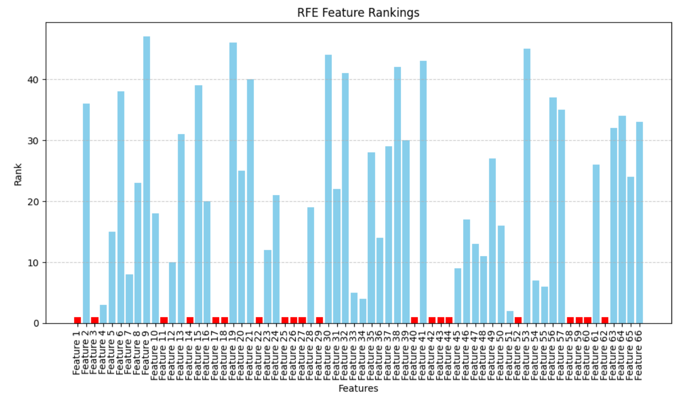

Fig. 6: RFE Feature Ranking

We employed optimization techniques to further enhance our feature selection process. Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving most of the important information. PCA identifies orthogonal axes, called principal components, that capture the maximum variance in the data. By selecting features associated with these principal components, PCA helps reduce the dimensionality of the dataset while retaining its essential information.

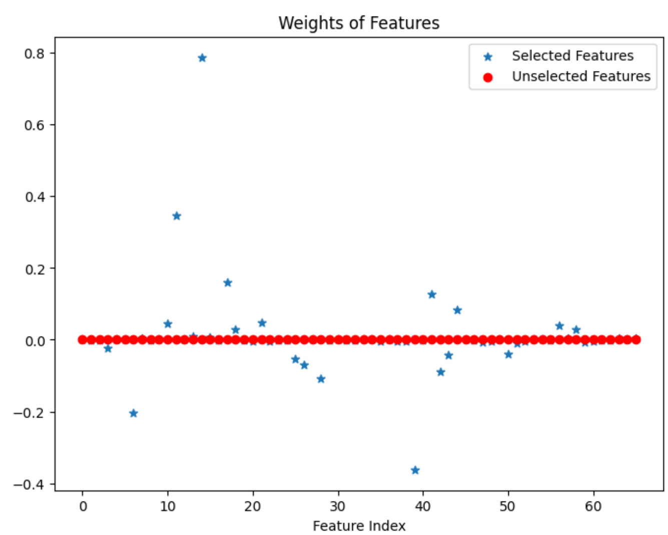

Fig. 7: Weights of Features

Particle Swarm Optimization (PSO) is another optimization technique we utilised. Inspired by the social behaviour of bird flocking or fish schooling, PSO is a population-based optimization algorithm. It iteratively updates a swarm of candidate solutions, known as particles, to identify the subset of features that optimise a predefined fitness function. PSO parameters, such as swarm size, inertia weight, and acceleration coefficients, play a crucial role in determining the algorithm's convergence behaviour and exploration-exploitation trade-off.

By employing these techniques in our feature selection process, we aimed to identify the most relevant features for our predictive modelling task, thereby improving the model's performance and interpret ability.

To identify the most impactful factor on the stress levels of engineering students, various analytical techniques were used to assess the importance of several features. These features include the student's passion for engineering, their role in internships, development of future skills, ability to manage stress, feelings of being overwhelmed by college, aspirations for higher studies, ability to deal with arguments, real-time agitation, feelings of pride, impact of recession, routine patterns, and self-care practices. The resulting table shows the importance of each feature as determined by the different techniques employed.

TABLE IV: Percentage of Important Features

Feature	Percentage
engg_passion	16%
Internship_role	16%
future_skill_development	13%
stress_management	13%
college_overhelmed	13%
higher_studies	12%
argument_deal	11%
realtime_agitate	10%
proud_feeling	10%

IV. Results and Discussion

A. Outcomes

In this study, a stacked ensemble model was employed to forecast the psychological well-being of engineering students. The ensemble comprised five base models: Logistic Regression (LR), K-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Support Vector Machines (SVM), and Naive Bayes. Five different meta-models were utilized: Support Vector Machines (SVM), Decision Tree Classifier, Logistic Regression, Multi-Layer Perception (MLP), and Gradient Descent. The table below summarizes the base model accuracies and stacking accuracies for each of the meta-Models:

TABLE V: Accuracy of Base Models and Meta Models

Models	SVM	DT	LR	MLP	GD
Logistic Regression	83.5%	83.5%	83.5%	83.3%	83.3%
K-Nearest Neighbours	83.6%	83.6%	83.6%	83.6%	83.6%
CART	68.4%	68.8%	68.8%	68.8%	68.8%
SVM	**90.7%**	**90.5%**	**90.5%**	**90.7%**	**90.7%**
Naive Bayes	82.4%	82.7%	82.7%	82.4%	82.4%
Stacking Acc.	**90.5%**	**87.0%**	**87.0%**	**90.8%**	**90.2%**

Note: DT is Decision Tree, LR is Logistic Regression, MLP is Multi-Layer Perceptron, GD is Gradient Descent, CART is Classification and Regression Trees.

Among the five meta-models evaluated, the Multi-Layer Perceptron (MLP) meta-model achieved the highest stacking accuracy of 90.8%, indicating its effectiveness in integrating the predictions of the base models. The Gradient Descent meta-model closely followed with a stacking accuracy of 90.2%. The Support Vector Machines (SVM) meta-model and base model consistently exhibited the highest individual accuracy of 90.7% across all meta-models. However, when combined with the MLP meta-model, the stacking accuracy further improved to 90.8%, surpassing the individual SVM model's performance. The Decision Tree Classifier and Logistic Regression meta-models both yielded a stacking accuracy of 87.0%, which, although lower than the top-performing meta-models, still demonstrated substantial improvements over the individual base models. Overall, the results highlight the efficacy of the stacked ensemble approach, particularly with the MLP and Gradient Descent meta-models, in accurately forecasting the psychological well-being of engineering students by effectively combining the predictions of multiple base models.

The table below presents the performance metrics for the Support Vector Machines (SVM), Classification and Regression Trees (CART), Logistic Regression (LR), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP) models used for classification

TABLE VI: Performance Evaluation of Classification Models

Model	Recall (%)	Precision (%)	Accuracy (%)	F1 Score (%)
SVM	96.22	96.22	96.00	96.23
CART	92.45	93.03	92.33	92.74
LR	96.22	96.22	96.00	96.22
GB	94.33	95.54	94.66	94.93
MLP	**96.22**	**96.83**	**96.33**	**96.52**

The MLP model achieved the highest performance, with an F-score of 96.52%, followed by the SVM and LR models (F-score: 96.23% and 96.22%, respectively). The GB model demonstrated good performance (F-score: 94.93%), while the CART model had the lowest performance (F-score: 92.74%). The high F-scores of the top-performing models indicate their effectiveness in accurately classifying instances related to forecasting the psychological well-being of engineering students.

B. Hyper-parameters

Like `n_neighbors` in KNN affects local pattern sensitivity, `max_iter` and `C` in logistic regression control optimization iterations and regularization strength, `kernel`, `C`, and `gamma` in SVM determine kernel type, regularization, and kernel coefficient respectively. For decision trees, `max_depth` limits tree depth and `min_samples_split` governs node splitting. Adjusting these hyperparameters tunes the model behavior for better performance.

- In this project, it explored various feature selection techniques, including ANOVA, Mutual Information (MI), Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and Particle Swarm Optimization (PSO), to optimize the performance of our stacked ensemble model.
- ANOVA and MI: Utilizing statistical techniques, it identified relevant features, achieving accuracies of 90.33% and 92.33% respectively.

TABLE VII: Table No 7: Hyperparameter Tuning

Classifier	Hyperparameter	Value
KNN	`n_neighbour`	5
LOGISTIC REGRESSION	`max_iter`	10000
	`C`	1.0
	`solver`	`lbfgs`
SVM	`kernel`	`rbf`
	`C`	1.0
	`gamma`	`scale`
DECISION TREE CLASSIFIER	`max_depth`	None
	`min_samples_split`	2

- RFE (Recursive Feature Elimination): Employing an iterative approach based on feature ranking, RFE yielded an accuracy of 90.33%. Interestingly, the selected features overlapped significantly with those from ANOVA and MI.
- PCA and PSO: holding optimization techniques, PCA and PSO demonstrated remarkable effectiveness, with accuracies of 92.33% and 96.33% respectively. These methods showcased their ability to discern the most appropriate features for model training, reinforcing the importance of optimization in enhancing performance.

TABLE VIII: Feature Selection Technique Evaluation

Feature Selection Technique	**Accuracy**
ANOVA	90.33%
Mutual Information	92.33%
RFE	90.33%
PCA	92.33%
PSO	96.33%

Across all feature selection techniques, certain features consistently stood out as highly influential. Notable examples include `stress_management`, `routine_pattern`, and `engg_passion`, underscoring their significance in predicting outcomes.
It utilized a heatmap to visualize pairwise correlations among the selected features. It helps identify strong positive and negative correlations, clusters of correlated features, and potential redundancies. This aids in identifying key features, detecting multicollinearity, and guiding feature engineering efforts. Most Impactful Features: Across all feature selection techniques, certain features consistently stood out as highly influential. Notable In the context of `engg_passion`, `internship_role`, `Future_skill_development`, `stress_management`, `College_overhelmed`, `higher_studies`, `argument_deal`, `Realtime_agitate`, `proud_feeling`, `recession_impact`, `routine_pattern`, and `selfcare` are important aspects to consider. underscoring their significance in predicting outcomes. We utilized a heatmap to visualize pairwise correlations among the selected features. It helps identify strong positive and negative correlations, clusters of correlated features, and potential redundancies. This aids in identifying key features, detecting multicollinearity, and guiding feature engineering efforts.

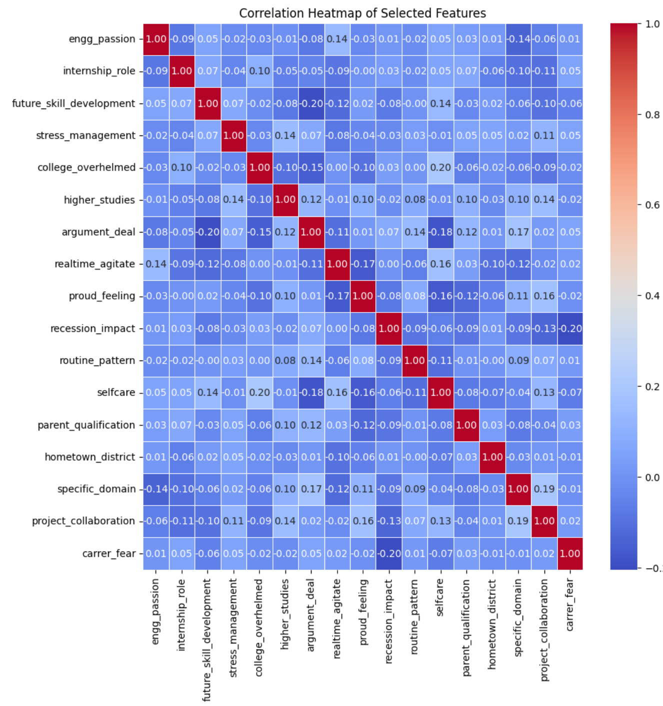

Fig. 8: Heatmap

C. *Confusion Matrix Analysis:*

The confusion matrix provides insight into the model's performance. It correctly classified 136 instances of Class 0 (Negative) and 153 instances of Class 1 (Positive). However, it misclassified 5 instances of Class 0 as Class 1 and 6 instances of Class 1 as Class 0, indicating areas for potential improvement in model robustness.

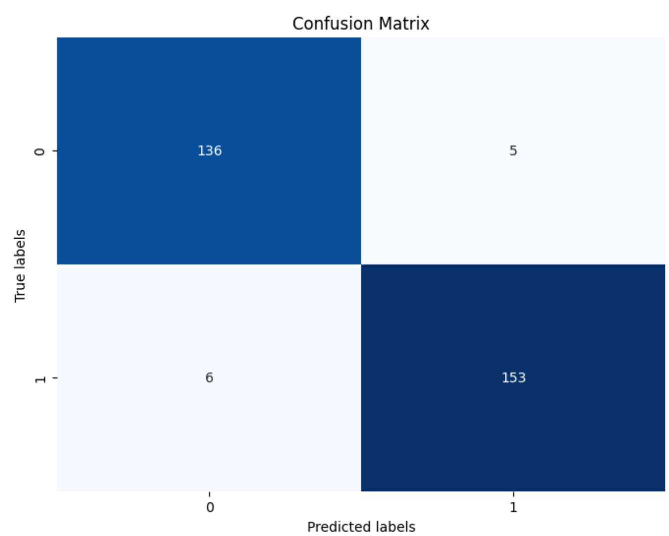

Fig. 9: Confusion Matrix

Overall, the combination of diverse feature selection methods and a stacked ensemble model enabled comprehensive analysis and robust predictive performance, laying a solid foundation for effective decision-making based on the dataset.

V. Conclusion

This project looked at how engineering students are doing mentally by studying different factors that affect their well-being. It collected data, cleaned it up, and used fancy math to build a model that predicts how students are feeling. By using smart techniques, it found important factors like stress management and routine habits. It also checked how well our model worked and saw where it could do better. This study explores how accurately it can understand stress in engineering students. Traditional methods, like assessments conducted by psychologists, may fall short and complicated when students are too stressed to provide accurate responses. Statistical methods also have limitations in accuracy. The research highlights the potential of advanced data analytics in identifying impactful stress factors. This approach offers practical advantages in proactively identifying and supporting stressed students, enabling targeted interventions and rehabilitation efforts to mitigate the adverse effects of stress on academic performance and overall well-being. Overall, this research helps us understand student mental health better and suggests ways to support them in college.

References

[1] Ahuja, R., Banga, A.: Mental stress detection in university students using machine learning algorithms. Procedia Computer Science **152**, 349–353 (2019), https://www.sciencedirect.com/science/article/pii/S1877050919306581

[2] Chung, J., Teo, J.: Mental health prediction using machine learning: taxonomy, applications, and challenges. Applied Computational Intelligence and Soft Computing **2022**, 1–19 (2022), https://api.semanticscholar.org/CorpusID:246001269

[3] D, A.: Stress level prediction of college students using ml. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (2023), https://api.semanticscholar.org/CorpusID:261929151

[4] Halde, R.R., Deshpande, A., Mahajan, A.: Psychology assisted prediction of academic performance using machine learning. In: 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). pp. 431–435. IEEE (2016), https://api.semanticscholar.org/CorpusID:18642275

[5] Manjunath, P., Shreya, P., Twinkle, S., Ashok, V., Sultana, S.: Predictive analysis of student stress level using machine learning. Int. J. Eng. Res. Technol **9**, 76–80 (2021)

[6] Marouf, A.A., Ashrafi, A.F., Ahmed, T., Emon, T.: A machine learning based approach for mapping personality traits and perceived stress scale of undergraduate students. International Journal of Modern Education & Computer Science **11**(8) (2019)

[7] Mutalib, S., et al.: Mental health prediction models using machine learning in higher education institution. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**(5), 1782–1792 (2021), https://turcomat.org/index.php/turkbilmat/article/view/2181

[8] Nison, P., Vuttipittayamongkol, P., Boonyapuk, P., Kemavuthanon, K.: A machine learning approach for depression screening in college students based on non-clinical information. 2023 International Conference On Cyber Management And Engineering (CyMaEn) pp. 413–417 (2023), https://api.semanticscholar.org/CorpusID:257260648

[9] Panthee, S., Rajkarnikar, S., Begum, R.: Career guidance system using machine learning. Journal of Advanced College of Engineering and Management (2023), https://api.semanticscholar.org/CorpusID:259744368

[10] Rahman, H.A., Kwicklis, M., Ottom, M.A., Amornsriwatanakul, A., Abdul-Mumin, K.H., Rosenberg, M., Dinov, I.D.: Prediction modeling of mental well-being using health behavior data of college students. Research Square (2022), https://api.semanticscholar.org/CorpusID:246591359

[11] Roy, M., Bhoi, A.K., Sharma, K.: Multimodal machine learning approaches for career prediction. 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) pp. 1–5 (2022), https://api.semanticscholar.org/CorpusID:257959960

[12] Shafiee, N.S.M., Mutalib, S.: Prediction of mental health problems among higher education student using machine learning. International Journal of Education and Management Engineering (IJEME) **10**(6), 1–9 (2020)

[13] Sharma, S., Pandey, S.K., Garg, K.: Machine learning for predictions in academics. International Journal of Recent Technology and Engineering (2020), https://api.semanticscholar.org/CorpusID:242741837

[14] Srividya, M., Mohanavalli, S., Bhalaji, N.: Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems **42**, 1–12 (2018)

[15] Sun, Y., Wu, H., Li, H., Fu, Y.: Machine learning-based mental health analysis and early warning for college student. 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C) pp. 569–578 (2021), https://api.semanticscholar.org/CorpusID:247858356

[16] Wang, C., Yang, T.: Research on mental health situations of college students based on machine learning algorithm. 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) pp. 1–6 (2023), https://api.semanticscholar.org/CorpusID:267125640

[17] Zhen, Y.: Prediction of college students' mental health based on status data. In: Other Conferences (2023), https://api.semanticscholar.org/CorpusID:25888063