### 1. Abstract

This study aims to predict depression among students using regression-based machine learning models. By leveraging a dataset containing academic, lifestyle, and demographic factors, the project explores relationships between predictors such as academic pressure, financial stress, sleep duration, and CGPA with depression outcomes. Models including Logistic Regression and Support Vector Machines (SVM) are implemented and compared based on their predictive performance. The results identify key risk factors for depression and provide insights into the role of data-driven approaches in student mental health monitoring.

### 2. Keywords

* Depression Prediction
* Regression Models
* Student Mental Health
* Machine Learning
* Predictive Analytics

### 3. Introduction

#### Background and Motivation

Depression among students is a critical issue affecting their academic performance, personal growth, and overall well-being. Research indicates that mental health problems are on the rise among college students due to factors such as academic pressure, financial stress, and lifestyle habits. Early identification of depression through data-driven methods can facilitate timely intervention and support.

Machine learning and regression-based models offer promising tools for predicting depression by identifying patterns and relationships in data. By analyzing a combination of demographic variables (e.g., age, gender), academic metrics (e.g., CGPA, study satisfaction), and lifestyle factors (e.g., sleep duration, financial stress), predictive models can help uncover significant risk factors for depression.

#### Research Objective

The goal of this project is to develop and evaluate regression-based models to predict depression among students. Specifically, the study will perform exploratory data analysis (EDA), build and compare regression models such as Logistic Regression and SVM, and identify key predictors of depression. The models will be evaluated using metrics such as accuracy, ROC-AUC, and other relevant measures.

#### Contribution

This study contributes to the growing body of research on student mental health by building interpretable regression models to predict depression outcomes, highlighting the importance of academic and lifestyle variables, and providing a framework for using data-driven methods to address mental health concerns. The findings aim to assist educators, mental health professionals, and policymakers in understanding the key factors contributing to depression and implementing targeted interventions.

### 4. Literature Review

#### Existing Approaches to Depression Prediction

Machine learning methods such as Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees, and Naive Bayes classifiers have been extensively applied to predict depression and anxiety among students. Logistic Regression is often favored for its simplicity and interpretability, making it a reliable baseline for binary classification tasks. Studies like those conducted by Sahu et al. (2023) on university students in Odisha revealed that hyperparameter-tuned logistic regression outperformed other models, achieving 92% accuracy​.

Naive Bayes classifiers, known for computational efficiency, have also demonstrated their utility in depression prediction. Priya et al. (2020) highlighted the model’s effectiveness in classifying depression and anxiety levels. Similarly, Ahmed et al. (2020) compared Naive Bayes with other supervised learning models like SVM and CatBoost, showing that while Naive Bayes is practical, ensemble methods often yield higher accuracy​​.

SVM models have shown robustness in handling high-dimensional and complex datasets. Hilbert et al. (2017) employed SVM for distinguishing generalized anxiety disorders from major depressive disorders, demonstrating its strength in separating overlapping mental health conditions when combined with features like MRI data and hormonal measures​.

#### Comparative Studies

The comparative analysis of machine learning models often shows trade-offs between simplicity and performance. Logistic Regression, while interpretable, may struggle with nonlinear relationships, whereas SVM and ensemble methods like Random Forest and Gradient Boosting Machines provide better performance in complex datasets. Studies by Sau and Bhakta (2017) illustrated Random Forest achieving 89% accuracy in predicting depression and anxiety among elderly patients​. However, these models often require more computational resources and extensive tuning.

#### Gaps in Existing Research

The existing literature predominantly emphasizes demographic variables like age and gender but often underrepresents behavioral and lifestyle factors such as sleep duration, dietary habits, and financial stress. Moreover, the generalizability of findings is limited by small sample sizes and region-specific datasets, as seen in studies focusing on Odisha or single university settings​.

#### Relevance to This Study

This study builds on these findings by incorporating a broader set of predictors, including academic pressure, financial stress, and dietary habits, which are often overlooked in previous research. By leveraging Logistic Regression and SVM models, it evaluates the balance between simplicity and robustness in predicting depression outcomes among students. The inclusion of one-hot encoding for categorical variables and advanced feature selection techniques ensures the models are well-suited to handle diverse predictors and offer meaningful insights.

### 5. Methodology

#### Data Description

The dataset comprises 27,901 student records, including both categorical and numerical variables. The target variable, Depression (binary: 0 = No, 1 = Yes), is predicted using a variety of academic, lifestyle, and demographic predictors. Key numerical variables include Age, CGPA, Academic Pressure, Work Pressure, Study Satisfaction, Job Satisfaction, Work/Study Hours, and Financial Stress. Categorical variables include Gender, Sleep Duration, Dietary Habits, Suicidal Thoughts, and Family History of Mental Illness. Variables such as Profession, City, ID, and Degree were excluded due to their limited relevance and imbalance in representation. To focus on a homogeneous group, only student records were retained, and other professions, which constituted a small minority, were excluded. Missing data in the Financial Stress variable (three records, approximately 0.01%) was addressed through mean imputation.

#### Data Preprocessing

Categorical variables were one-hot encoded to ensure compatibility with machine learning models, and numerical variables such as Age and CGPA were standardized using z-scores to prevent scale differences from influencing the models. The dataset was then split into training (75%) and testing (25%) subsets, ensuring that the class distribution of the target variable remained consistent across both splits. Feature selection was performed to retain relevant predictors, excluding variables such as ID, which was deemed non-informative, and Profession and City, which exhibited high categorical imbalance. These preprocessing steps ensured the dataset was ready for regression-based modeling.

#### Model Development

Two models were developed and tested for predicting depression: Logistic Regression (LR) and Support Vector Machines (SVM). Logistic Regression was chosen as a baseline model for its simplicity, interpretability, and efficiency in binary classification tasks. In contrast, SVM was selected for its ability to model nonlinear relationships using the radial basis function (RBF) kernel, making it well-suited for capturing complex data patterns. Both models were trained using 10-fold cross-validation to enhance robustness and prevent overfitting. For SVM, hyperparameter tuning was conducted to optimize the sigma and C values, ensuring peak performance.

#### Model Evaluation

The models were evaluated using a comprehensive set of metrics, including accuracy, sensitivity, specificity, and ROC-AUC. Accuracy measured overall prediction correctness, while sensitivity evaluated the ability to correctly identify depressed students. Specificity measured the ability to identify non-depressed students, and ROC-AUC assessed the models' capacity to distinguish between classes. All modeling and evaluation processes were conducted using the caret package in R, which provided tools for cross-validation, feature preprocessing, and hyperparameter optimization.

### 6. Experimentation and Results

The dataset preparation involved excluding variables with limited relevance, imputing missing values, and encoding categorical features. The refined dataset was split into training and testing sets, where the training set was used to build the models and the testing set was reserved for evaluation.

For the Logistic Regression model, the training phase achieved an ROC score of 0.92 with sensitivity of 0.79 and specificity of 0.88. On the test set, the model achieved an accuracy of 84.38% (95% CI: 83.51% to 85.23%), with a balanced accuracy of 83.58%. Sensitivity was 0.79, while specificity was 0.88, demonstrating strong overall predictive performance.

The Support Vector Machines (SVM) model was trained using optimized hyperparameters, with sigma = 0.031 and cost (C) = 0.25. During training, the SVM model achieved an ROC score of 0.92, sensitivity of 0.79, and specificity of 0.89, with low standard deviations across the cross-validation folds. On the test set, the SVM model achieved an accuracy of 84.21% (95% CI: 83.33% to 85.06%), with a Kappa value of 0.6713 and a balanced accuracy of 83.33%. Sensitivity was 0.78, and specificity was 0.88, reflecting strong classification performance.

Additionally, the Generalized Linear Model (GLM) was evaluated, achieving consistent results across both training and testing phases. On the training set, the GLM recorded an ROC of 0.923 with a sensitivity of 0.886 and specificity of 0.796. Similarly, the test set results mirrored these metrics with an ROC of 0.923, sensitivity of 0.886, and specificity of 0.796. These findings indicate the model's robustness and its ability to generalize well across different subsets of data.

Both models demonstrated reliable and consistent results, with comparable accuracy, sensitivity, and specificity. The Logistic Regression model slightly outperformed SVM in terms of sensitivity, identifying a higher proportion of depressed students. However, the SVM model provided similar overall performance with marginally better specificity, indicating a stronger ability to correctly identify non-depressed students. These results suggest that both models are effective tools for predicting depression outcomes, with Logistic Regression offering simplicity and interpretability, while SVM demonstrates robustness in handling more complex relationships within the data.

### 7. Discussion

### Discussion

The results of this study align with findings in the existing literature that highlight the effectiveness of Logistic Regression and Support Vector Machines for predicting depression outcomes. Logistic Regression, with its simplicity and strong performance in identifying depressed students (sensitivity of 0.79), mirrors conclusions drawn by prior studies, where high accuracy was reported in a similar context. However, the model's performance can be limited when relationships in the data become more complex.

The SVM model demonstrated comparable overall performance with marginally higher specificity (0.88), highlighting its ability to correctly identify non-depressed students. This result aligns with prior research showing that SVM models excel in high-dimensional and complex datasets due to their robust handling of non-linear relationships. While SVM requires careful parameter tuning, as evidenced by the optimization of sigma and cost parameters in this study, it remains a reliable choice for binary classification tasks.

The Generalized Linear Model (GLM) further demonstrated robust predictive performance, achieving the highest sensitivity (0.886) among all evaluated models. This ensures that a greater proportion of depressed students were correctly identified, a critical aspect in mental health applications where early detection is paramount. However, GLM exhibited slightly lower specificity (0.796) compared to the other models, which may lead to a marginally higher rate of false positives.

Compared to ensemble methods such as Random Forests or Gradient Boosting Machines, as discussed in the reviewed literature, both Logistic Regression and SVM provide a balance between accuracy and interpretability. Ensemble models, while achieving higher accuracies in some studies, often lack the transparency needed for understanding key predictors. This trade-off underscores the value of Logistic Regression and SVM in contexts where model interpretability is essential for actionable insights.

The inclusion of diverse predictors such as academic pressure, financial stress, and sleep duration distinguishes this study from prior work that often relied on limited demographic variables. While Naive Bayes models discussed in the literature have highlighted the importance of feature selection, this study demonstrates that incorporating a broad range of academic and lifestyle factors improves predictive performance without overfitting.

Overall, Logistic Regression emerges as a robust and interpretable tool for depression prediction, particularly when sensitivity is prioritized. SVM offers slight improvements in specificity, making it suitable for applications requiring precise identification of non-depressed cases. GLM, with its superior sensitivity, is a critical model for scenarios emphasizing the identification of at-risk students. Future work could explore integrating ensemble models or additional predictors, such as behavioral and social data, to further enhance prediction accuracy and generalizability.

### 8. Conclusion

This study demonstrates the utility of regression-based machine learning models for predicting depression among students, with Logistic Regression, SVM, and GLM all exhibiting strong predictive performance. Logistic Regression proved to be a robust, interpretable model, excelling in sensitivity and ensuring the identification of a high proportion of depressed students. The SVM model provided comparable results, with slightly improved specificity, making it an effective tool for reducing false positives. Meanwhile, the GLM demonstrated exceptional sensitivity, highlighting its suitability for applications where identifying at-risk individuals is critical.

By incorporating diverse predictors such as academic pressure, financial stress, and sleep duration, this study bridges gaps identified in prior research, offering a more comprehensive approach to depression prediction. The results underscore the importance of balancing model complexity and interpretability, especially in mental health contexts where actionable insights are essential for early intervention.

Future work could focus on incorporating additional predictors, such as behavioral and social factors, and exploring ensemble methods to further enhance prediction accuracy. The findings of this study provide valuable insights for educators, mental health professionals, and policymakers, paving the way for data-driven approaches to address the growing mental health challenges among students.

### References

1. Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., & Torres-Ceclén, C. (2024). Machine Learning Models to Classify and Predict Depression in College Students. International Journal of Interactive Mobile Technologies (iJIM), 18(14), pp. 148–163. https://doi.org/10.3991/ijim.v18i14.48669
2. Priya, S., Garg, R., & Sharma, K. (2023). Machine Learning Models to Classify and Predict Depression. International Journal of Interactive Mobile Technologies (iJIM), 18(14), 148-163. Retrieved from https://doi.org/10.3991/ijim.v18i14.48669.

### Future Work

This study establishes a foundation for using regression-based machine learning models to predict depression among students, but there are several opportunities for future research:

1. Incorporation of Additional Predictors: Future models could include behavioral and social data, such as interaction patterns on digital platforms, extracurricular activities, and social network analysis, to gain deeper insights into depression risk factors.
2. Exploration of Advanced Models: While Logistic Regression, SVM, and GLM proved effective, advanced models such as Random Forests, Gradient Boosting Machines, and Neural Networks could be explored to enhance prediction accuracy and address complex data relationships.
3. Longitudinal Data Analysis: Incorporating longitudinal data could provide insights into how depression evolves over time and the temporal influence of various predictors.
4. Cross-Cultural Validation: Extending the study to datasets from different cultural and geographical contexts would help validate the models and adapt them to diverse populations.
5. Intervention-Driven Modeling: Developing models that not only predict depression but also suggest personalized interventions based on key predictors would make the approach more actionable for mental health professionals.
6. Integration with Real-Time Monitoring: Integrating predictive models into real-time systems, such as mobile apps or online platforms, could facilitate continuous monitoring and early detection of depression symptoms among students.