Project Name

211401037 Mustafa Güneyli
211401032 İrem Öztürk
211401023 Mesut Özkan

A report submitted as part of the requirements for the
CEN411 Data Mining Course
Department of Computer Engineering
Recep Tayyip Erdoğan University
Rize, Türkiye

December 2025

Instructor: Dr.Büşra ÇALMAZ

Abstract

This project investigates the relationship between students' stress levels and their academic performance using data mining techniques. The dataset includes important features such as sleep duration, study hours, family support, and stress level, all of which are recognized in the literature as key indicators of academic outcomes. After collecting and preprocessing the dataset, exploratory data analysis was conducted to better understand the distribution, correlation patterns, and possible outliers among the variables.

Supervised machine learning models, including Logistic Regression and Decision Tree, were applied to predict academic performance based on stress-related attributes. These models were selected because they offer both predictive capability and interpretability. The predictive results showed that stress level, sleep quality, and family support played a significant role in determining student outcomes, aligning with findings from previous research.

In addition to supervised learning, K-Means clustering was implemented to discover hidden patterns among students with similar stress characteristics. Clustering results revealed distinct student groups, such as high-stress/low-performance and low-stress/high-performance clusters. These insights highlight the importance of early identification of at-risk students and the potential for educational interventions such as counseling, mentoring, or workload adjustments.

Overall, the results indicate that stress has a notable negative impact on academic performance and can be effectively analyzed using data mining methods. The combination of supervised and unsupervised learning provided both predictive insights and student profile segmentation. The findings of this project may guide future studies and contribute toward developing strategies that improve student well-being and learning outcomes.

Contents

Chapter 1

Introduction

1.1 Problem Definition and Data Understanding

1.1.1 Problem Definition

Academic stress is a growing concern that negatively affects students' academic performance and well-being. Factors such as heavy academic workload, irregular study habits, insufficient sleep, and limited family support contribute to elevated stress levels. Students experiencing high stress often demonstrate reduced concentration, lower academic achievement, and decreased engagement in learning activities. The primary objective of this project is to examine how stress-related factors influence academic performance and to develop predictive models for identifying students at academic risk. Additionally, the study aims to uncover patterns among students with similar stress characteristics, providing insights that may support educational interventions to enhance both academic success and mental well-being.

1.1.2 Dataset Description

The dataset used in this study includes variables related to students' stress and lifestyle, such as sleep duration, study hours, family support, and perceived stress level. These attributes are widely recognized in the literature as important indicators of academic performance. The dataset contains both numerical and categorical variables, making it suitable for supervised and unsupervised machine learning methods. Prior to model development, data preprocessing was performed to address missing values, normalize feature scales, and examine correlations. This process ensured the accuracy, reliability, and interpretability of the subsequent analyses.

1.2 Conclusion

In conclusion, this chapter establishes the foundation of the study by defining the research problem and describing the dataset. Understanding the relationship between stress-related factors and academic performance enables more effective prediction and profiling of student outcomes. The following chapters present the methodology, modeling approaches, clustering analysis, and evaluation of results in detail.

Chapter 2

Literature Review

2.1 Stress Factors in Student Life

Student stress is shaped by a combination of academic and personal factors, with academic workload and time management difficulties being the most prominent contributors. Continuous pressure to succeed in examinations often leads students to sacrifice rest and leisure, increasing the risk of chronic stress and burnout. Previous studies report that excessive academic demands are strongly associated with elevated stress levels and declining psychological well-being among students [1][2]. Sleep duration and family support further influence students' stress experiences. Insufficient sleep negatively affects cognitive functioning and emotional regulation, while supportive family environments promote resilience and effective stress coping. In contrast, lack of family support intensifies anxiety and disengagement [3]. These factors frequently interact, highlighting the need for a holistic assessment of stress determinants.

2.2 Relationship Between Stress and Academic Performance

A substantial body of research indicates a generally negative relationship between student stress and academic performance. High stress levels impair concentration, memory, and learning efficiency, resulting in lower academic achievement and reduced engagement [1]. Empirical findings consistently show that increased stress correlates with declines in GPA, examination results, and course completion rates. However, this relationship is not strictly linear. According to the Yerkes–Dodson law, moderate levels of stress may enhance short-term performance, whereas excessive or prolonged stress significantly impairs academic functioning [4]. Long-term unmanaged stress is closely linked to anxiety, depression, and academic burnout, further reinforcing its negative impact on student success [5].

2.3 Predictive Modeling of Academic Performance Using Machine Learning

Supervised machine learning approaches have gained increasing importance in predicting academic outcomes due to their ability to model complex and nonlinear relationships. Algorithms such as logistic regression and decision trees are commonly employed to classify student performance and identify individuals at academic risk using demographic, behavioral, and psychological features. Prior research emphasizes that interpretable models, particularly logistic regression and decision trees, offer valuable insights into the role of stress-related and support-based variables in academic performance [2]. These findings support the selection of similar models in the present study to balance predictive accuracy and interpretability.

2.4 Clustering in Educational Data for Student Profile Discovery

Unsupervised learning techniques, especially k-means clustering, are widely applied to uncover hidden student profiles within educational datasets. By grouping students with similar psychological and behavioral characteristics, clustering enables the identification of distinct stress and performance patterns. Studies demonstrate that clustering methods can reveal vulnerable student groups, such as those experiencing high stress and low academic achievement, thereby supporting targeted academic and psychological interventions [3]. Such approaches provide insights that extend beyond the capabilities of supervised models alone.

2.5 Synthesis of Insights and Relevance to the Present Study

The reviewed literature consistently shows that student stress is influenced by academic workload, sleep quality, and family support, all of which significantly affect academic performance. Unfavorable conditions increase stress and reduce learning outcomes, whereas balanced workloads and supportive environments promote academic success [1][2].

These findings justify the inclusion of stress-related and lifestyle variables in the present dataset and support the combined use of supervised learning and clustering techniques.

Chapter 3

Design

3.1 Data Collection

The dataset used in this study was obtained from the Kaggle platform and is based on a survey conducted among university students. The dataset contains information related to students' demographic characteristics, academic background, and mental health conditions. In total, the dataset consists of 101 instances and 11 attributes. The collected attributes include age, gender, field of study, year of study, cumulative grade point average (CGPA), and several psychological indicators such as depression, anxiety, and panic attacks. These variables are directly relevant to the aim of this project, which is to analyze how stress-related factors affect students' academic performance. The dataset was obtained from a single source; therefore, no data merging or integration process was required. It is considered appropriate for this problem because it explicitly includes mental health indicators that can be used to model student stress levels alongside an academic performance measure.

3.2 Data Preprocessing

This section describes all preprocessing steps applied to the dataset prior to exploratory data analysis and model building. Data preprocessing was a crucial step to ensure data quality, consistency, and suitability for machine learning algorithms.

3.2.1 Handling Missing Values and Duplicates

An initial examination of the dataset revealed a small number of missing values in the *Age* attribute. To preserve the dataset size and avoid information loss, the missing value was replaced using a central tendency approach. After this step, the dataset contained no missing values. Duplicate record analysis was also conducted, and no duplicated observations were detected. Therefore, no duplicate removal was required. After handling missing values and duplicates, the final dataset consisted of 101 instances and 11 attributes.

3.2.2 Noise Removal and Data Cleaning

During the data cleaning phase, inconsistencies were identified in categorical variables, particularly in the CGPA attribute. Variations in category formatting, such as differences in spacing and label representation, were standardized to ensure consistent category definitions. This cleaning process prevented incorrect grouping and misclassification during target variable construction and subsequent analyses. No additional noise smoothing techniques were required, as the dataset consisted primarily of categorical and binary variables.

3.2.3 Normalization and Standardization

Some machine learning algorithms used in this study are sensitive to the scale of input features. Therefore, normalization and standardization techniques were applied where appropriate. For Logistic Regression and K-Means clustering, numerical features were standardized using z-score normalization to ensure that all variables contributed equally to distance calculations and optimization processes. For the Naïve Bayes model, Min–Max normalization was applied to continuous features to map values into a bounded range suitable for probabilistic modeling. All scaling procedures were implemented within machine learning pipelines to reduce the risk of data leakage.

3.2.4 Feature Engineering

The original dataset did not include a direct numerical representation of students' stress levels. To address this limitation, additional features were engineered to explicitly capture stress-related information in a structured and interpretable manner.

First, a composite variable termed *stress score* was constructed by aggregating the binary indicators of depression, anxiety, and panic attacks. This score ranges from 0 to 3, with higher values indicating the presence of a greater number of stress-related psychological symptoms. The main purpose of this feature was to provide a unified and continuous representation of overall stress severity.

Second, a binary feature named *high stress without treatment* was derived to identify students experiencing elevated stress levels (defined as a stress score greater than or equal to 2) who had not sought professional psychological support. This feature was specifically designed to capture the potential risk associated with unmanaged or untreated stress conditions.

Overall, these engineered features enriched the dataset by transforming individual psychological indicators into more informative and analytically useful variables. This process enhanced model interpretability and facilitated a more meaningful examination of the relationship between stress, mental health, and academic performance.

3.2.5 Feature Selection Method

Feature selection was conducted using a filter-based approach informed by correlation analysis and domain knowledge. The correlation analysis revealed a strong association between the engineered *stress score* and the *high stress without treatment* variable, which was expected given their related construction.

To mitigate the risk of multicollinearity and avoid redundant information, these highly correlated features were not included simultaneously within the same supervised learning models. Instead, alternative model configurations were evaluated using one of these features at a time.

The final feature set was selected to balance predictive relevance and interpretability, and it consisted of demographic attributes and stress-related variables most aligned with the research objectives. This approach ensured stable model behavior while preserving meaningful psychological insights.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the statistical properties of the dataset, examine relationships among variables, and identify patterns prior to model building. This section presents summary statistics, visual analyses, class distribution, correlation patterns, and attribute skewness, along with interpretations of the key findings.

3.3.1 Summary Statistics

Summary statistics such as mean, median, variance, and range were calculated for numerical and engineered features. The engineered stress score had a mean value of approximately 1.01, with values ranging from 0 to 3. This indicates that while most students experienced low to moderate stress levels, a smaller subset of students reported high stress. Correlation analysis between stress-related variables and academic performance revealed very weak linear relationships, suggesting that academic success is influenced by multiple interacting factors rather than stress indicators alone.

3.3.2 Class Distribution and Imbalance Analysis

The distribution of the target variable (`high_cgpa`) was examined to assess potential class imbalance. The dataset contained 53 students (52.48%) labeled as low or medium academic performance and 48 students (47.52%) labeled as high academic performance. This relatively balanced distribution indicates that severe class imbalance is not present.

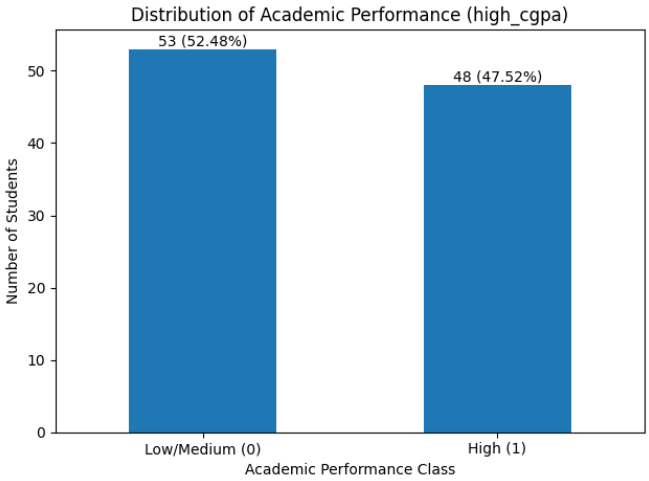

Figure 3.1: Distribution of academic performance classes (high_cgpa).

3.3.3 Stress Score Distribution

The distribution of the engineered stress score was analyzed to examine overall stress patterns among students. As illustrated in Figure 3.2, the majority of students had stress scores of 0 or 1, indicating low stress levels. However, approximately 28% of students had stress scores of 2 or 3, representing moderate to high stress conditions. This result highlights the presence of a meaningful subgroup of students experiencing elevated stress, justifying further analysis of stress-related characteristics.

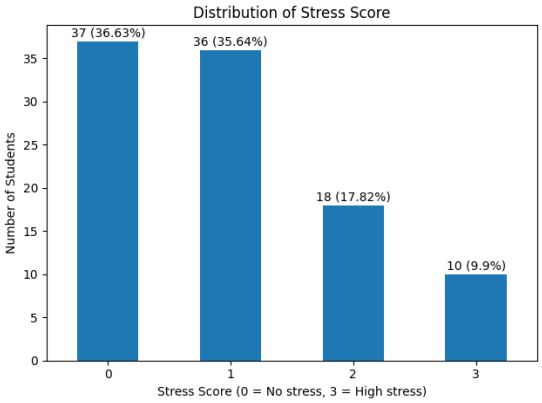

Figure 3.2: Distribution of stress score values among students.

3.3.4 Stress Score and Academic Performance

To explore the relationship between stress levels and academic performance, a boxplot comparison was performed. The median stress score appeared similar across academic performance groups. However, the low and medium academic performance group exhibited a wider spread of stress values and a higher concentration of moderate to high

stress scores. This observation suggests that stress alone does not directly determine academic success, but elevated stress levels are more frequently observed among students with lower academic performance.

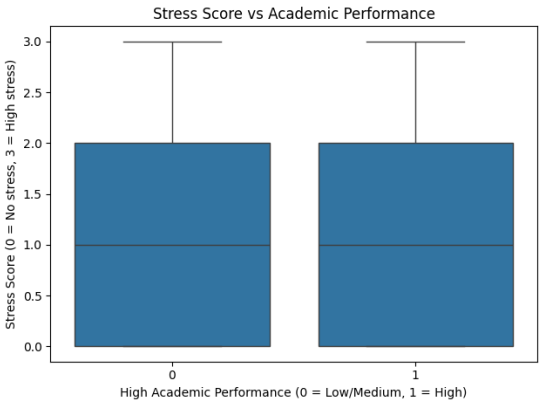

Figure 3.3: Stress score distribution across academic performance classes.

3.3.5 Risk Group Analysis

Students were further classified into risk groups using the `high_stress_no_treatment` feature. A comparison of academic performance across these groups showed that students experiencing high stress without seeking professional treatment had a slightly lower proportion of high academic performance compared to the non-risk group. Although the difference is modest, this trend suggests that unmanaged psychological stress may negatively influence academic outcomes.

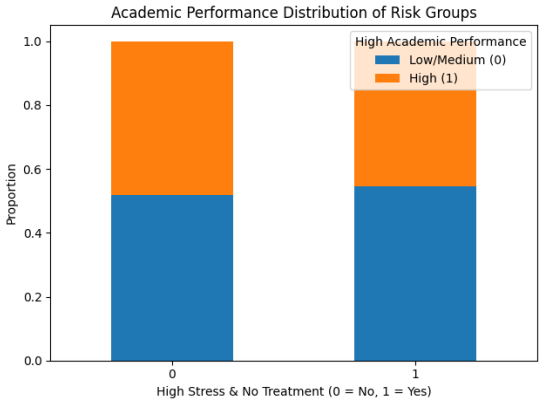

Figure 3.4: Academic performance distribution across risk and non-risk groups.

3.3.6 Correlation Analysis

Correlation analysis was conducted to examine linear relationships among selected numerical and binary variables. The correlation between stress score and academic performance was found to be very weak, indicating that stress-related variables alone are

insufficient predictors of academic success. A strong positive correlation was observed between stress score and the high stress without treatment variable, which is expected due to the feature derivation process.

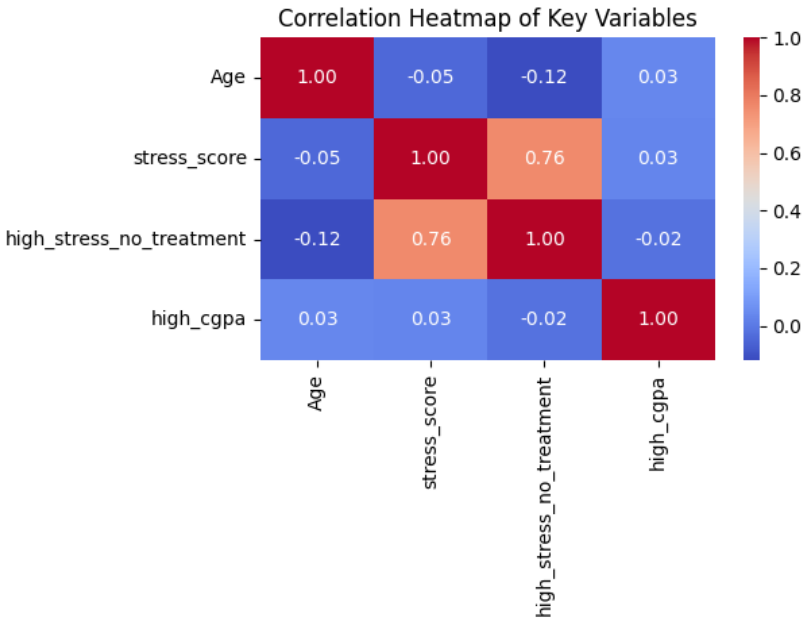

Figure 3.5: Correlation heatmap of selected numerical and binary variables.

3.3.7 Attribute Skewness

Attribute skewness was analyzed to assess the symmetry of key numerical variables. The age variable exhibited mild positive skewness, while the stress score showed moderate positive skewness, indicating that most students reported low stress levels with fewer students experiencing high stress. This observation is visually supported by the histogram and kernel density estimation plots shown in Figure 3.6.

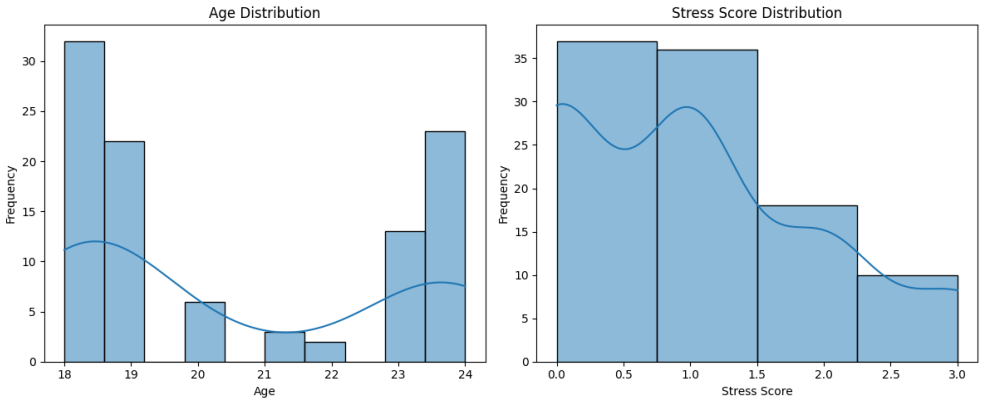

Figure 3.6: Histogram and KDE plots illustrating the distributions of Age and stress score.

Table 3.1: Skewness values of selected attributes.

Attribute	Skewness
Age	0.39
stress_score	0.64

3.4 Model Building / Data Mining

This section presents the data mining and machine learning models implemented to analyze the relationship between students' mental health indicators and academic performance. Both supervised and unsupervised learning approaches were applied. For supervised learning, multiple classification models were trained and evaluated using the same feature set and preprocessing pipeline to ensure a fair and consistent comparison.

3.4.1 Supervised Methods

Supervised learning models were used to predict high academic performance (`high_cgpa`) based on demographic and psychological features, including age, stress score, depression, anxiety, and panic attack indicators. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices.

Logistic Regression

Logistic Regression was implemented as a baseline classifier due to its simplicity and interpretability, using a pipeline with median imputation, feature standardization, and class-balanced training. The model achieved the highest overall performance, with an accuracy of 0.6667 and a precision of 0.8000 for the high academic performance class. Despite a relatively low recall (0.4000), the high precision indicates fewer false positive predictions. As shown in Figure 3.7, the model correctly classified most low or medium performance students while misclassifying some high-performing students.

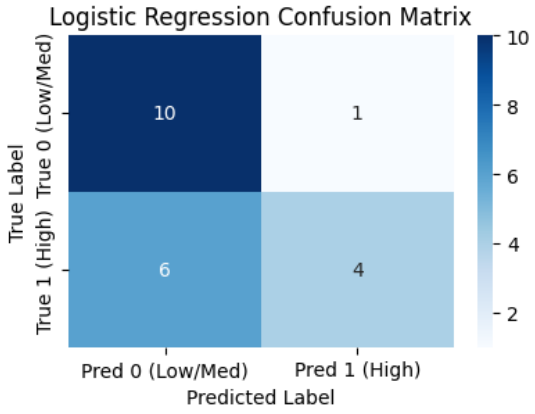

Figure 3.7: Confusion matrix of the Logistic Regression model

Decision Tree

The Decision Tree classifier was applied to model non-linear relationships between mental health indicators and academic performance. Unlike linear models, Decision Trees generate explicit decision rules, enhancing interpretability. The model achieved an accuracy of 0.5238, with balanced precision and recall values (0.5000). Figure 3.8 shows that the model correctly classified half of the high-performing students. Although its predictive performance was moderate, the Decision Tree provided valuable insights into feature interactions, particularly the role of depression status, stress score, and age.

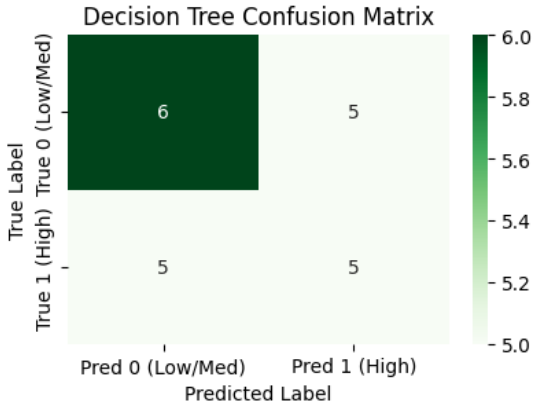

Figure 3.8: Confusion matrix of the Decision Tree classifier

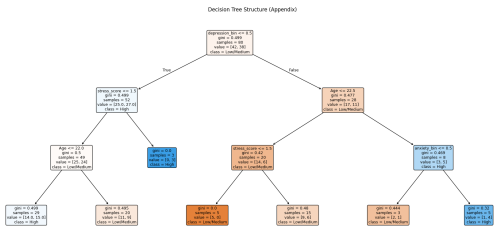

Figure 3.9: Learned structure of the Decision Tree model

Bernoulli Naive Bayes

Bernoulli Naive Bayes was implemented due to the binary nature of most mental health indicators. Continuous variables were scaled to the [0,1] range to meet model assumptions. The model achieved an accuracy of 0.5714. While precision for the high academic performance class was moderate, recall remained low (0.4000), indicating difficulty in identifying all high-performing students. The confusion matrix in Figure 3.10 shows that the model favored predicting the low or medium performance class.

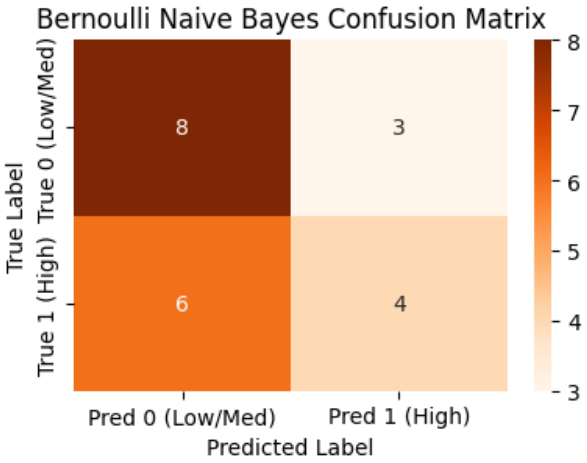

Figure 3.10: Confusion matrix of the Bernoulli Naive Bayes classifier

k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) algorithm was employed as a distance-based classifier. Feature standardization was applied prior to training. The k-NN model exhibited the lowest performance among supervised classifiers, achieving an accuracy of 0.4762. The recall value for high academic performance was particularly low (0.3000), as illustrated in Figure 3.11. This result can be attributed to the small dataset size and sparse neighborhood structure.

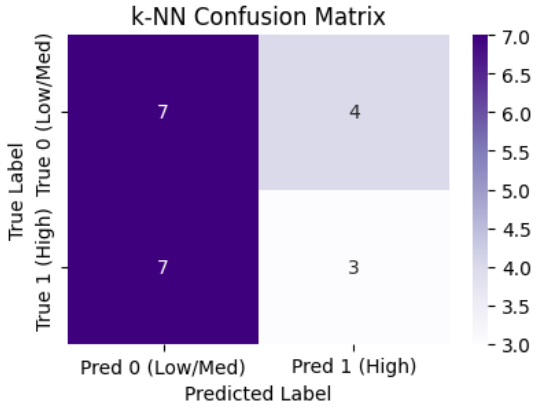

Figure 3.11: Confusion matrix of the k-Nearest Neighbors classifier

Random Forest

Random Forest was applied as an ensemble learning method that combines multiple decision trees to improve robustness. Balanced class weights were used to address class imbalance.

The model achieved an accuracy of 0.4762 and did not significantly outperform simpler classifiers, as shown in Figure 3.12. Feature importance analysis (Figure 3.13) revealed that age was the most influential predictor, followed by stress score and anxiety-related variables.

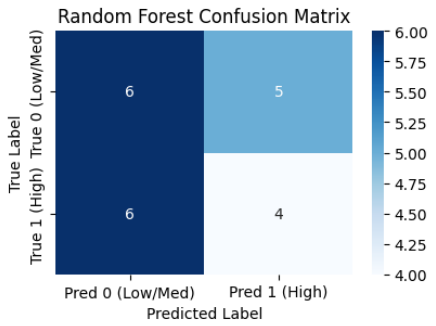

Figure 3.12: Confusion matrix of the Random Forest classifier

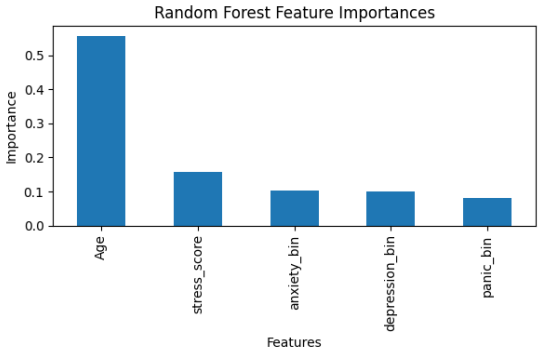

Figure 3.13: Feature importance scores obtained from the Random Forest model

Model Comparison

A comparative evaluation of the supervised learning models was conducted using accuracy, precision, recall, and F1-score metrics. All models were trained and evaluated using identical preprocessing steps and feature sets.

Table 3.2: Performance comparison of supervised learning models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.6667	0.8000	0.4000	0.5333
Decision Tree	0.5238	0.5000	0.5000	0.5000
Naive Bayes	0.5714	0.5714	0.4000	0.4706
k-Nearest Neighbors	0.4762	0.4286	0.3000	0.3529
Random Forest	0.4762	0.4444	0.4000	0.4211

Logistic Regression emerged as the best-performing supervised model in terms of accuracy and precision. However, all models exhibited relatively low recall for high academic performance, suggesting that mental health indicators alone may not be sufficient to fully explain academic success.

3.4.2 Unsupervised Methods

Unsupervised learning techniques were applied to explore latent psychological patterns among students without using academic performance labels.

K-Means Clustering

K-Means clustering was applied to identify distinct psychological stress profiles among students using the standardized features *stress_score*, *depression_bin*, *anxiety_bin*, and *panic_bin*. Academic performance variables were intentionally excluded to ensure that the clustering process was entirely unsupervised. Prior to clustering, all features were standardized using the StandardScaler to prevent scale dominance in distance-based calculations. The optimal number of clusters was determined using the Elbow Method, which evaluates within-cluster sum of squares for different values of k. As shown in Figure 3.14, a noticeable reduction in inertia was observed up to $k = 3$, after which the improvement became marginal. Therefore, three clusters were selected for the final K-Means model.

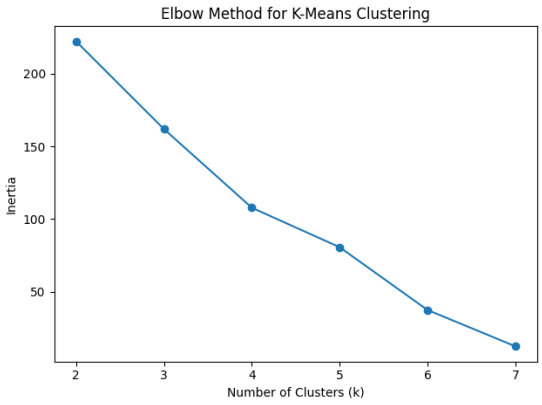

Figure 3.14: Elbow method for selecting the optimal number of clusters in K-Means

After applying K-Means with three clusters, students were assigned to distinct groups representing different stress profiles. The distribution of students across clusters is reported in Table 3.3, showing a balanced separation without extreme cluster dominance.

Table 3.3: Distribution of students across K-Means clusters

Cluster	**Number of Students**
Cluster 0	11
Cluster 1	4
Cluster 2	6

To evaluate clustering quality, the Silhouette Score was computed, yielding a value of 0.36. This score indicates a moderate but acceptable level of cluster separation, which is reasonable given the limited dataset size and the binary nature of several features. Cluster characteristics were further examined by computing the mean values of each feature within clusters. Figure 3.15 presents a heatmap of cluster profiles, revealing

clear psychological distinctions among groups. One cluster exhibits high stress scores and elevated anxiety and panic indicators, representing a high-stress group. Another cluster shows consistently low values across all features, corresponding to a low-stress or psychologically healthy group. The remaining cluster displays moderate stress levels with a notable tendency toward panic attacks.

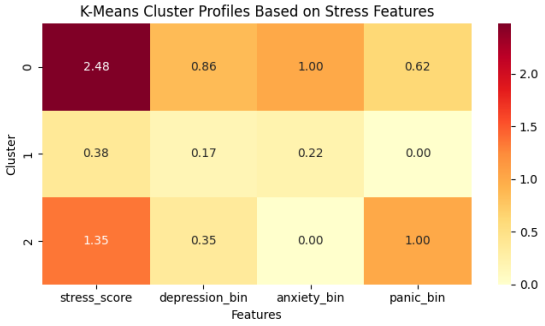

Figure 3.15: K-Means cluster profiles based on stress-related features

Overall, K-Means clustering successfully identified meaningful and interpretable psychological stress profiles among students. These findings provide valuable complementary insights to the supervised learning analysis and demonstrate the usefulness of unsupervised methods in uncovering latent patterns within mental health data.

Hierarchical Clustering (Agglomerative)

Hierarchical clustering was applied as an alternative unsupervised method to identify stress-related student profiles. Agglomerative clustering with Ward linkage was used, as it minimizes within-cluster variance and produces well-structured clusters for standardized data.

A dendrogram was generated to examine the hierarchical structure of the data and to determine the appropriate number of clusters. As shown in Figure 3.16, a clear separation was observed, supporting the selection of three clusters. This choice is consistent with the K-Means clustering results and allows direct comparison between methods.

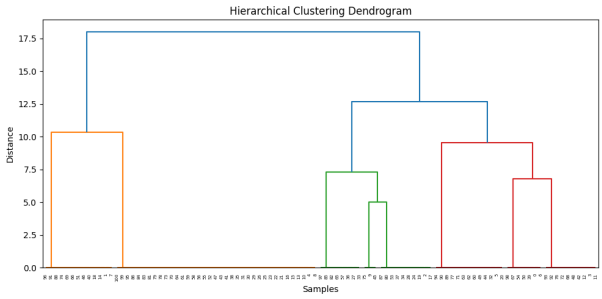

Figure 3.16: Hierarchical clustering dendrogram using Ward linkage

The clustering quality was evaluated using the Silhouette Score, which reached 0.528, indicating good cluster separation. Cluster profiles were analyzed by examining the mean values of stress-related features within each group. The resulting clusters represent low-stress, moderate-stress, and high-stress student profiles, as illustrated in Figure 3.17.

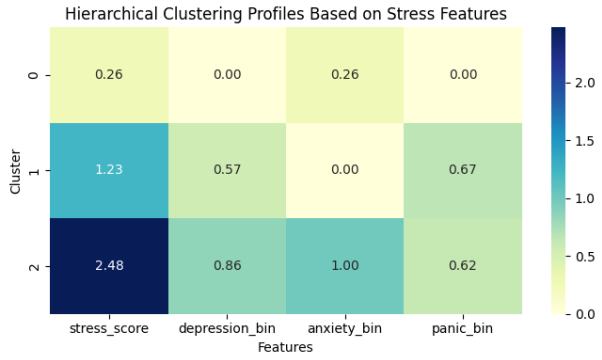

Figure 3.17: Hierarchical clustering profiles based on stress-related features

Overall, hierarchical clustering produced well-separated and interpretable stress profiles and demonstrated stronger clustering performance compared to K-Means for this dataset.

DBSCAN

DBSCAN was applied as a density-based clustering method to identify stress-related student groups and potential outliers. Unlike K-Means and hierarchical clustering, it does not require a predefined number of clusters and can label sparse observations as noise. The method produced multiple small clusters with a very low noise ratio. Although the silhouette score was high, this result was driven by the formation of small, homogeneous clusters rather than well-separated global structures. As illustrated in Figure 3.18, DBSCAN captured fine-grained local variations in stress-related features, but the resulting clusters were fragmented and difficult to interpret at a global level.

Overall, DBSCAN was less effective than K-Means and hierarchical clustering in identifying meaningful and interpretable psychological stress profiles for this dataset.

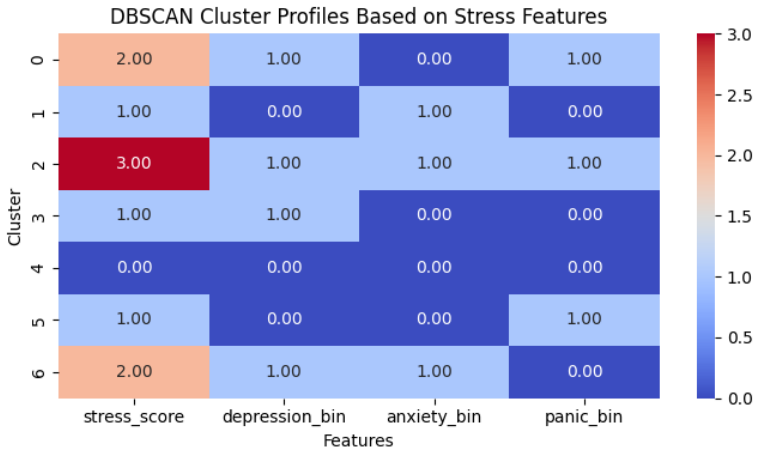

Figure 3.18: DBSCAN cluster profiles based on stress-related features

3.4.2 Unsupervised Model Comparison The unsupervised learning methods were compared based on cluster structure, interpretability, silhouette score, and the ability to detect noise. A summary of the comparison is presented in Table 3.4.

Table 3.4: Comparison of unsupervised clustering methods

Method	Clusters	Silhouette Score	Noise Detected	Interpretability
K-Means	3	0.36	No	Moderate
Hierarchical	3	0.53	No	High
DBSCAN	Multiple	1.00	Yes	Low

Among the evaluated unsupervised methods, hierarchical clustering demonstrated the strongest overall performance by achieving the highest silhouette score while producing clearly interpretable and well-separated stress profiles. K-Means clustering also identified meaningful groups but exhibited weaker cluster separation. In contrast, DBSCAN generated numerous small clusters with limited global interpretability, despite its ability to detect a small number of outliers. These findings indicate that hierarchical clustering is the most suitable unsupervised method for uncovering stress-related psychological patterns in this dataset.

3.5 Conclusion

This study examined the relationship between university students' mental health indicators and academic performance using data mining and machine learning techniques.

A structured workflow was followed, including data preprocessing, feature engineering, exploratory data analysis, supervised classification, and unsupervised clustering, to obtain both predictive and interpretative insights. Exploratory data analysis revealed that stress-related variables exhibit meaningful patterns across student groups, although their direct linear relationship with academic performance is weak. The distribution of the engineered stress score (Figure 3.2) showed that while most students reported low stress levels, a notable subgroup experienced moderate to high stress. The comparison of stress score across academic performance groups (Figure 3.3) indicated that elevated stress levels were more frequently observed among students with lower academic performance, suggesting an indirect influence of stress on academic outcomes. Further evidence was provided by the risk group analysis based on the `high_stress_no_treatment` feature. As shown in Figure 3.4, students experiencing high stress without professional support demonstrated a slightly lower proportion of high academic performance. Although this difference was modest, it supports the assumption that unmanaged psychological stress may negatively affect academic success. Correlation analysis (Figure 3.5) confirmed that stress-related variables alone are insufficient predictors of academic performance, highlighting the multifactorial nature of academic achievement. Among the supervised learning models, Logistic Regression achieved the best overall performance in terms of accuracy and precision (Table 3.2). Its confusion matrix (Figure 3.7) indicates fewer false positive predictions compared to other classifiers, explaining its higher precision. However, recall values remained relatively low across all models, indicating limitations in identifying all high-performing students using mental health indicators alone. Feature importance analysis from the Random Forest model (Figure 3.13) further emphasized the relevance of age and stress-related features.

Unsupervised learning methods complemented the supervised analysis by revealing latent psychological stress profiles. K-Means and hierarchical clustering identified distinct and interpretable stress-related groups (Figures 3.15 and 3.17), with hierarchical clustering demonstrating stronger cluster separation as indicated by its higher Silhouette score (Table 3.4). In contrast, DBSCAN produced fragmented clusters with limited global interpretability (Figure 3.18).

Overall, the findings suggest that mental health indicators are meaningfully associated with academic performance but do not fully explain academic success on their own. The combined use of supervised and unsupervised methods enabled a more comprehensive understanding of student stress patterns and their potential academic implications. Future work may incorporate additional academic, behavioral, and socio-demographic variables to improve predictive performance and deepen insight into the relationship between mental health and academic outcomes.

Chapter 4

Evaluation & Testing

4.1 Model Evaluation

This section presents the evaluation results of the supervised and unsupervised learning models implemented in this study. The models were assessed using standard performance metrics in order to ensure an objective and consistent comparison. For classification tasks, Accuracy, Precision, Recall, F1-score, and ROC/AUC metrics were employed. For clustering tasks, the Silhouette coefficient was used to evaluate cluster separation quality.

Classification Performance

The performance of the supervised classification models is summarized in Table 4.1. Logistic Regression achieved the highest overall accuracy (0.6667) and precision (0.8000), indicating reliable identification of high academic performance cases. However, recall values across all models remained relatively low, suggesting that predicting high academic performance based solely on mental health indicators is challenging.

Tree-based and probabilistic models demonstrated moderate performance, while k-Nearest Neighbors and Random Forest yielded lower accuracy and F1-scores. These results highlight the limitations of complex models when applied to small-scale datasets with limited feature diversity.

Table 4.1: Supervised model performance comparison

Model	Accuracy	Precision	Recall	F1-score	Best(Acc)
Logistic Regression	0.6667	0.8000	0.4000	0.5333	✓
Decision Tree	0.5238	0.5000	0.5000	0.5000	
Naive Bayes	0.5714	0.5714	0.4000	0.4706	
k-Nearest Neighbors	0.4762	0.4286	0.3000	0.3529	
Random Forest	0.4762	0.4444	0.4000	0.4211	

ROC/AUC Analysis

To further evaluate the discriminative capability of the classification models, Receiver Operating Characteristic (ROC) analysis was conducted. Figure 4.1 presents the ROC curve of the Logistic Regression model.

The Logistic Regression classifier achieved a ROC–AUC score of 0.727, indicating a moderate ability to distinguish between high and low academic performance classes. This result confirms that the model performs substantially better than random classification, while also suggesting that additional features may be required to improve predictive performance.

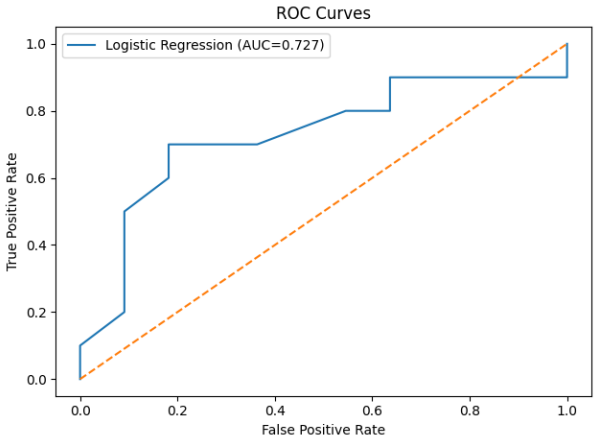

Figure 4.1: ROC curve of the Logistic Regression model

Clustering Evaluation

The evaluation results of the unsupervised learning methods are presented in Table 4.2. Hierarchical clustering achieved the highest Silhouette coefficient (0.528), indicating stronger cluster separation compared to K-Means clustering (0.360). DBSCAN identified multiple small clusters and detected a small proportion of noise points; however, its cluster structure was fragmented, resulting in lower overall interpretability.

Table 4.2: Unsupervised method evaluation results

Method	**Clusters**	**Silhouette**	**Noise Detected**	**Interpretability**
K-Means	3	0.360	No	Moderate
Hierarchical	3	0.528	No	High
DBSCAN	Multiple	1.000	Yes (3%)	Low

Overall, hierarchical clustering demonstrated the most robust clustering performance by producing well-separated and interpretable psychological stress profiles. These findings suggest that hierarchical methods are particularly suitable for uncovering latent mental health patterns in small-scale student datasets.

4.2 Interpretation and Reporting

The interpretation of the results was supported by visual analyses presented in the previous sections. In particular, confusion matrices, feature importance plots, and clustering visualizations were used to better understand model behavior and to justify the selection of the best-performing methods.

The confusion matrices presented for the supervised models (Figures 3.7–3.12) provide detailed insight into classification errors. Among these, the Logistic Regression confusion matrix (Figure 3.7) shows a lower number of false positive predictions compared to other models, explaining its higher precision value. This figure was therefore central in identifying Logistic Regression as the most reliable supervised model. In contrast, the confusion matrices of k-Nearest Neighbors and Random Forest indicate a higher rate of misclassification, particularly false negatives, which aligns with their lower recall values.

Feature importance analysis further supports these findings. The Random Forest feature importance plot (Figure 3.13) highlights age and stress score as the most influential predictors of academic performance. Although Random Forest did not achieve the highest predictive performance, this figure was valuable for identifying which mental health indicators contribute most to the classification task and for interpreting the supervised learning results beyond accuracy metrics.

For unsupervised learning, cluster visualizations played a key role in interpreting latent psychological patterns. The K-Means cluster profile heatmap (Figure 3.15) illustrates clear differences between low-, moderate-, and high-stress student groups, demonstrating that meaningful psychological profiles can be identified without using academic performance labels. Similarly, the hierarchical clustering dendrogram and cluster profile visualization (Figures 3.16 and 3.17) show well-separated and interpretable clusters, which justifies the selection of hierarchical clustering as the most effective unsupervised method.

In contrast, the DBSCAN cluster profile heatmap (Figure 3.18) reveals fragmented and highly specific clusters. Although DBSCAN achieved a high silhouette score, this figure indicates that the method captures localized patterns rather than globally interpretable stress profiles. This visual evidence explains why DBSCAN was considered less suitable for this dataset despite its strong numerical clustering metric.

Overall, the combined use of performance metrics and visual analyses provides a comprehensive interpretation of the results. The figures confirm that logistic regression and hierarchical clustering offer the most reliable and interpretable solutions for supervised and unsupervised analysis, respectively.

Chapter 5

Conclusion

This chapter summarizes the main findings of the study, highlights its key contributions, and discusses possible directions for future research. The conclusions are drawn by jointly interpreting quantitative performance metrics and visual analyses presented in the previous chapters.

5.1 Conclusions

The primary objective of this study was to investigate the relationship between students' mental health indicators and academic performance using data mining and machine learning techniques. To address this objective, both supervised and unsupervised learning approaches were implemented and evaluated.

The supervised learning results indicate that mental health-related variables, including stress level, anxiety, depression, panic attacks, and age, are associated with academic performance; however, their predictive capability is limited when considered in isolation. As shown by the classification performance metrics and confusion matrices (Figures 3.7–3.12), Logistic Regression achieved the highest accuracy and precision among the evaluated models. The corresponding confusion matrix (Figure 3.7) demonstrates a lower number of false positive predictions, which explains the model's superior precision and supports its selection as the most reliable supervised classifier in this study.

Despite these results, the confusion matrices also reveal that all supervised models exhibit relatively low recall values, indicating difficulty in identifying all high-performing students. This limitation suggests that academic success cannot be fully explained by mental health indicators alone and that additional explanatory variables are required.

Unsupervised learning methods provided complementary insights by uncovering latent psychological stress profiles among students. The K-Means cluster profile visualization (Figure 3.15) illustrates clear differences between low-, moderate-, and high-stress groups, demonstrating the presence of meaningful psychological patterns in the data. Hierarchical clustering further improved upon this result by producing well-separated and highly interpretable clusters, as evidenced by the dendrogram and cluster profile visualizations (Figures 3.16 and 3.17). These figures justify the selection of hierarchical clustering as the most effective unsupervised method for this dataset.

In contrast, DBSCAN produced fragmented and highly localized clusters, as shown in Figure 3.18. Although DBSCAN achieved a high silhouette score, the corresponding cluster profiles indicate limited global interpretability. This visual evidence explains why DBSCAN was less suitable for identifying broad psychological stress patterns in the context of this study.

Overall, the combined analysis of numerical metrics and visual results demonstrates that Logistic Regression and hierarchical clustering provide the most reliable and interpretable outcomes for supervised and unsupervised analysis, respectively. The study highlights the importance of using multiple analytical perspectives to better understand the complex relationship between mental health and academic performance.

5.2 Future Work

Several directions for future research can be derived from the findings and limitations of this study. First, expanding the dataset to include a larger and more diverse student population would improve model generalizability and reduce sensitivity to class imbalance. Incorporating additional academic, behavioral, and socio-demographic features—such as attendance records, study habits, course difficulty, and socioeconomic background—may significantly enhance predictive performance and recall.

Second, future work may explore more advanced modeling approaches, including ensemble learning, gradient boosting techniques, and deep learning models, to capture more complex and non-linear relationships between mental health indicators and academic performance. Alternative preprocessing strategies and feature engineering methods could also be investigated to improve overall model balance.

Finally, longitudinal analysis using time-series mental health and academic data could provide deeper insights into how changes in psychological well-being influence academic outcomes over time. Such extensions would contribute to a more comprehensive and actionable understanding of student well-being and performance.

Chapter 6

Team Contributions

- **Mustafa Güneyli:** Responsible for the preliminary stages of the project, including **Chapter 1 (Introduction)** and **Chapter 2 (Background and Related Work)**. These contributions involved defining the research problem, outlining the motivation and objectives of the study, reviewing relevant literature, and establishing the theoretical foundations required for the analysis.
- **İrem Öztürk:** Responsible for the core analytical components of the project, including **Chapter 3 (Design)** and **Chapter 4 (Evaluation & Testing)**. This work covered data preprocessing, feature engineering, exploratory data analysis, implementation of supervised and unsupervised learning models, model evaluation using multiple performance metrics, interpretation of results, and preparation of figures and tables for the report.
- **Mesut Özkan:** Responsible for **Chapter 5 (Conclusion and Future Work)**. This contribution focused on summarizing the overall findings of the project, discussing key insights derived from the analysis, identifying limitations, and proposing directions for future research and improvement.

All team members collaboratively reviewed the final report, validated experimental results, and contributed to the preparation of the project presentation. This collaborative process ensured consistency across chapters and collective responsibility for the overall quality of the project outcomes.

Chapter 7

Code and Reproducibility Notes

7.1 Source Code

All code in this project was implemented in Python and organized in a modular structure covering data preprocessing, exploratory data analysis, feature engineering, supervised and unsupervised learning, and model evaluation.

The complete source code is available at the following GitHub repository:

- **GitHub Repository:** Student Stress and Academic Performance

The repository includes Python scripts (`.py`) and Jupyter notebooks (`.ipynb`) required to reproduce all experiments.

7.2 Execution Instructions

To reproduce the results, follow these steps:

1. Install Python (version 3.8 or higher).
2. Install required libraries:

   ```
   pip install pandas numpy scikit-learn matplotlib seaborn
   ```

3. Clone the GitHub repository.
4. Run the preprocessing scripts, followed by supervised and unsupervised model scripts.

7.3 Dataset Access

The dataset used in this study was obtained from Kaggle:

- **Dataset Source:** Student Mental Health Dataset

7.4 Reproducibility Statement

All preprocessing steps, model configurations, and evaluation procedures are fully documented and implemented in the provided codebase. Random seeds were fixed where applicable to ensure reproducibility.

Bibliography

[1] Sanchita Deb, Esben Strodl, and Hong Sun. "Academic stress, parental pressure, anxiety and mental health among Indian high school students". In: *International Journal of Psychology and Behavioral Sciences* 5.1 (2015), pp. 26–34.

[2] K. J. Reddy, K. R. Menon, and A. Thattil. "Academic stress and its sources among university students". In: *Biomedical and Pharmacology Journal* 11.1 (2018), pp. 531–537. DOI: 10.13005/BPJ/1404.

[3] Sunil Kumar and Jitender Bhukar. "Stress level and coping strategies of college students". In: *Journal of Physical Education and Sports Management* 4.1 (2013), pp. 5–11.

[4] Robert M. Yerkes and John D. Dodson. "The relation of strength of stimulus to rapidity of habit-formation". In: *Journal of Comparative Neurology and Psychology* 18.5 (1908), pp. 459–482.

[5] Cristina Mazza et al. "A nationwide survey of psychological distress among Italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors". In: *International Journal of Environmental Research and Public Health* 17.9 (2020), p. 3165. DOI: 10.3390/ijerph17093165.