

COMPARATIVE ANALYSIS OF CLASSIFICATION MODEL PERFORMANCE IN PREDICTING STRESS LEVELS IN STUDENTS

Ana Đokić¹ Hana Stefanović² Dragana Dudić³

Abstract: This paper presents a comparative analysis of the performance of classification models in predicting stress levels in students, with a particular focus on the correlation between variables such as study hours and sleep hours and stress levels. The correlation between these variables is established as the foundation for building and evaluating predictive models. Classification models such as logistic regression and decision tree were employed in the study, and their results were analyzed using metrics such as accuracy, precision, recall, and F1-score. This work contributes to the understanding of the application of machine learning in stress research within an educational context, offering insights into the advantages and challenges associated with various approaches.

Key words: Machine learning, classification, decision tree, logistic regression, stress, students

1. INTRODUCTION

Stress is one of the key factors influencing students' academic achievements, emotional states, and social interactions.

Stress among students represents one of the main factors affecting their academic performance, emotional well-being, and social interactions. Given the widespread presence of stress among students [1], it is important to understand the parameters that cause it and develop effective models for its identification and prediction. Parameters such as study hours, sleep hours, physical activity, and social interactions can have a significant impact on stress levels, and understanding their interrelationship may help in finding strategies to reduce stress among students [2].

This study utilized various machine learning models [3] to analyze and predict the impact of stress on students. Initially, a correlation analysis between different factors and stress levels was performed to identify the parameters that have the greatest influence on predicting stress. Based on this analysis, classification models such as logistic regression and decision tree [4] were applied to compare their performance in predicting stress levels.

This research aims to compare the effectiveness of different classification models in predicting stress while analyzing how correlations between factors such as study hours and sleep hours influence their accuracy.

2. MATERIALS AND METHODS

2.1. Dataset and Preprocessing

This study used a dataset [5] containing information about students, such as their academic and lifestyle habits, including study hours, sleep hours, physical activity, and social activities. The data underwent preprocessing [6] to ensure optimal preparation for model creation. Stress levels among students were classified into three categories: low, moderate, and high. The dataset was used for training and testing various classification models, to predict stress levels based on study hours, sleep hours, and other relevant factors.

The Python programming language was used for data analysis, enabling the use of necessary libraries such as pandas (data processing) [7], scikit-learn (model implementation) [8], as well as matplotlib [9] and seaborn [10] (data visualization) for preprocessing, model creation, and evaluation. All implementations were conducted in the free virtual environment Anaconda.

As part of the preprocessing, in addition to checking for missing and duplicate values, the ID column was removed as it is not related to the behaviors or characteristics under analysis. The

¹M.Sc, Information Technology School, Belgrade, email: ana.djokic@its.edu.rs

²PhD, School of Electrical and Computer Engineering, Belgrade, email: hana@viser.edu.rs

³PhD, Informational Technology School, Belgrade, email: dragana.dudic@its.edu.rs

Stress_Level column was encoded into numerical values to enable the application of machine learning algorithms, which require numerical input data for efficient analysis and prediction of stress levels. The values in the GPA column were scaled to a specific range, which helps standardize the data and ensures that attributes with different ranges do not dominate or disproportionately affect the performance of machine learning algorithms. The data were split into a training set and a test set, with 80% of the data used for training, and the remaining 20% used for testing in order to evaluate model performance.

Figure 1 shows information about the dataset, including column names, number of rows, columns, and data types, after preprocessing.

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Study_Hours_Per_Day                  2000 non-null   float64
1   Extracurricular_Hours_Per_Day       2000 non-null   float64
2   Sleep_Hours_Per_Day                 2000 non-null   float64
3   Social_Hours_Per_Day                 2000 non-null   float64
4   Physical_Activity_Hours_Per_Day      2000 non-null   float64
5   GPA                                  2000 non-null   float64
6   Stress_Level                         2000 non-null   object
7   Stress_Level_Num                     2000 non-null   int64
dtypes: float64(6), int64(1), object(1)
```

Figure 1 – Data set information of given data set, Source: Author's paper

2.2. Correlation Between the Given Variables and Stress Levels

To illustrate the relationship between various numerical variables in the dataset, a correlation matrix was created. In this case, its purpose is to explain how different factors (such as Study_Hours_Per_Day, Sleep_Hours_Per_Day, GPA, and others) are related to stress levels and to each other, in order to facilitate further analysis and data modeling. A visual representation of the interrelationship between the variables is shown in Figure 2.

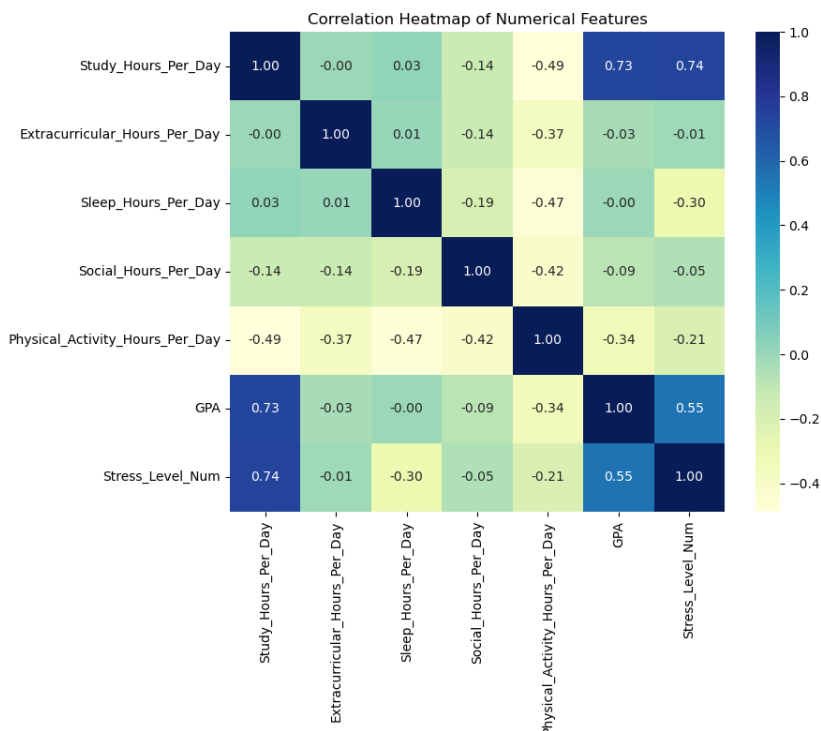


Figure 2 – Correlation Heatmap, Source: Author's paper

Based on the results shown in Figure 2, it can be concluded that the most prominent factors influencing stress in students are the number of study hours and sleep hours. Therefore, the strong positive correlation between the number of study hours per day and stress levels (value 0.74) suggests that as study hours increase, stress levels in students also increase. This is expected, as students who spend more time studying are often under greater pressure due to academic demands and deadlines. On the other hand, the moderately negative correlation between sleep hours and stress levels (value -0.3) implies that students who sleep more tend to have lower stress levels, while sleep deprivation can increase stress, and adequate sleep has a calming effect. This highlights the importance of time management in developing strategies for stress reduction, such as efficient study planning, breaks during study sessions, and promoting healthy habits to improve academic performance and general well-being. Some physical and social activities, also have some impact but are less dominant. It is interesting to note the correlation value between time spent on extracurricular activities and stress levels (value -0.01). This weak negative correlation suggests that students who spend more time on extracurricular activities do not experience significantly lower stress levels, or that this time is not sufficient to alleviate stress arising from academic responsibilities. In other words, while extracurricular activities may have a positive impact on relaxation, their effect on stress reduction may be limited, especially if students are already overwhelmed by study-related obligations.

3. CLASSIFICATION MODELS

3.1. Results and Analysis

Based on the previously presented correlation matrix, relevant variables were selected for model creation, as these variables are expected to have the greatest impact on model accuracy. Two classification models were selected for analysis: logistic regression and decision tree. Model evaluation was carried out using metrics such as accuracy, precision, recall and F1-score, as well as the confusion matrix, which shows how the model performs across different classes (low, moderate and high stress levels).

3.1.1. Logistic Regression

Logistic regression is a classification model used to model the probability of certain events. This model has a simple implementation and performs very well on linearly separable classes[4]. The results of the code for the logistic regression model created for this study are shown in Figure 3:

	precision	recall	f1-score	support
1	0.85	0.90	0.87	67
2	0.82	0.74	0.78	127
3	0.87	0.90	0.89	206

Figure 3 – Classification Report for Logistic Regression, Source: Author's paper

Based on the results shown in Figure 3, it can be concluded that the application of the logistic regression model achieved an accuracy of 85%, meaning that the classification of the majority of samples in the test set was successful. Precision and recall were generally balanced, but there were some variations among the classes.

For Class 1 (low stress) and Class 3 (high stress), the results were exceptionally good, with a recall of 0.90, indicating that the model is very effective at identifying these stress levels. However, Class 2 (moderate stress) had a lower recall (0.74), suggesting that the model did not always succeed in accurately identifying students with moderate stress. The F1-score also indicated a similar trend, with values between 0.78 and 0.89 for all classes, which suggests a good balance between precision and recall, with Class 2 having the lowest F1-score (0.78).

The graphical representation of the confusion matrix for the logistic regression model is shown in Figure 4.

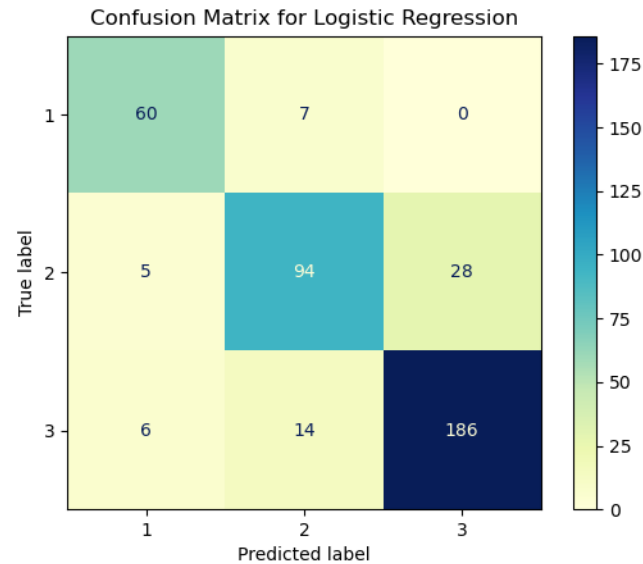


Figure 4 – Confusion Matrix for Logistic Regression, Source: Author's paper

Based on the visual representation in Figure 4, it can be concluded that the model is very good at recognizing high stress levels (Class 3), as it has a high value on the diagonal (186) and a low number of misclassified samples (6 for Class 1 and 14 for Class 2). The model performs somewhat weaker in classifying moderate stress levels (Class 2), as it has a relatively large number of misclassified samples (28 to Class 3 and 5 to Class 1). The model's recognition of low stress levels (Class 1) is weaker, with only 60 correct classifications and up to 7 misclassifications of moderate stress.

This matrix allows for a detailed analysis of the model's performance, as well as identifying where the model is making errors. Low stress levels (Class 1) are poorly recognized, while high and moderate stress levels are better identified. Using this information, the model can be optimized to improve performance, especially in recognizing low stress levels.

3.1.2. Decision Tree

Decision trees are classification and regression models that create a hierarchical decision structure. They are simple to implement and provide clear insights into the decision-making process. Decision trees are valuable and flexible for solving a wide range of problems, as they can handle both numerical and categorical data, while not requiring significant data preprocessing[4].

The results of the decision tree model code developed for this paper are presented in Figure 5:

	precision	recall	f1-score	support
1	1.00	1.00	1.00	67
2	1.00	1.00	1.00	127
3	1.00	1.00	1.00	206

Figure 5 – Classification Report for Decision Tree, Source: Author's paper

Based on the results shown in Figure 5, it can be concluded that the decision tree model achieved perfect results with an accuracy of 100%, precision, recall, and F1-score of 1.00 for all three classes (low, moderate, and high stress).

These results indicate that the model made no classification errors and successfully identified all three classes with complete precision and recall. This suggests that the decision tree model is highly effective for predicting stress levels among students.

The graphical view of the confusion matrix for the decision tree model is shown in Figure 6.

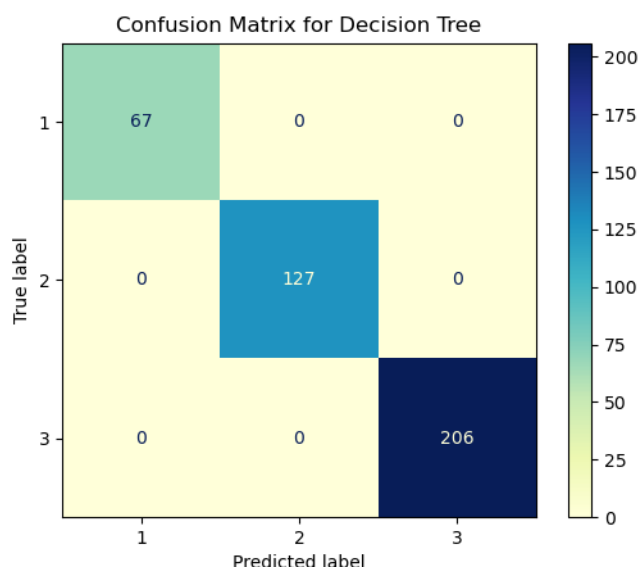


Figure 6 – Confusion Matrix for Decision Tree, Source: Author's paper

Based on the visual representation in Figure 6, it can be concluded that the confusion matrix clearly shows that the model achieved perfect performance. Given that all the numbers are on the diagonals (True Positives), this means that the model correctly classified all instances in the test set. Therefore, there were no false positives (False Positives) or false negatives (False Negatives), indicating that the model made no classification errors.

3.2. Comparative Analysis of the Logistic Regression and Decision Tree Models

Based on the comparative analysis of the key metrics of the logistic regression and decision tree models, the following conclusions can be drawn:

- The decision tree model achieved a perfect accuracy of 100%, meaning it classified all instances correctly without errors. In this regard, the decision tree proved to be the superior model in terms of accuracy. On the other hand, logistic regression had a lower accuracy of 85%, indicating that the model made some classification errors, although it still achieved relatively good results.
- Logistic regression exhibited precision values ranging from 0.82 (for class 2 – moderate stress) to 0.87 (for class 3 – high stress), which indicates how many of the predicted positive instances were correct. The decision tree, however, demonstrated perfect precision for all classes, meaning it classified each class without any prediction errors.
- Logistic regression showed similar recall values for classes 1 and 3 (0.90), but its recall for class 2 (moderate stress) was lower (0.74). This indicates that the model did not always succeed in recognizing all instances of class 2, which can be critical as a lower recall suggests a higher number of false negatives. In contrast, the decision tree achieved a perfect recall of 1.00 for all classes, meaning it successfully detected all instances.
- Logistic regression displayed solid F1-scores, but the values varied across classes: from 0.78 for class 2 to 0.89 for class 3. The lower F1-score for class 2 suggests that the model struggled with accurately detecting this stress level. The decision tree, however, achieved a perfect F1-score of 1.00 for all classes, indicating an ideal balance between precision and recall for all stress levels.
- Although the number of instances in class 2 (moderate stress) and class 1 (low stress) is smaller compared to class 3 (high stress), these numbers are not directly related to the model performance. It is important to note that the decision tree model achieved perfect performance even for the classes with fewer instances, while logistic regression had weaker performance for class 2.

The decision tree model achieved perfect performance in all key metrics (precision, recall, F1-score), with 100% accuracy. This means that the model classified all instances correctly, making it seemingly very reliable for predicting stress levels. However, there is a possibility that the model is overfitted and achieved excellent results solely because it was too tailored to the specifics of the training set. Given the similarity of the test data to the training set, it is likely, that the model's excellent performance may falter when tested on real-world, unseen data.

Logistic regression, which is simpler and faster to implement, showed lower performance, particularly in classifying class 2 (moderate stress). While the model achieved solid results, it was unable to recognize all instances of class 2, reducing the overall accuracy and recall for this class.

4. CONCLUSION

The models used for classification include logistic regression and decision tree, both of which were trained on the training dataset and evaluated on the test dataset. Evaluation methods involve accuracy, precision, recall, and F1-score metrics. Although the decision tree can achieve perfect results under specific conditions, logistic regression is likely the better choice when stability, the ability to generalize to unseen data, and balanced performance across classes are required.

Future research directions would include expanding the dataset to cover various factors that influence student stress, such as social and emotional factors, and exploring the impact of demographic characteristics. While logistic regression and decision tree provided good results, exploring other models, such as random forest, could improve model accuracy as they utilize multiple trees to make predictions, thus reducing the likelihood of overfitting. With 100 trees ($n_{\text{estimators}}=100$), random forest is typically more robust and less prone to overfitting compared to individual decision trees. Additionally, expanding the analysis of correlations among all relevant factors (such as study hours, sleep hours, GPA, and other socio-economic factors) may help identify key variables affecting stress levels. These insights could be valuable for further development of the model and its capabilities.

It is important to note that the dataset used in this study was analyzed primarily for model implementation and does not reflect real-life data specific to the student population in this study. For further research, a more comprehensive analysis could be conducted using actual data collected from students in real-world conditions, allowing for more accurate predictions and insights that are tailored to the specific characteristics and needs of the target student group. This would enable a more robust understanding of the factors influencing student stress levels in real academic environments.

5. REFERENCES

- [1] Akanpaadgi, E., Binpinbu, F.: *The impact of stress on students' academic performance*, Eureka Journal of Educational Research 2(1): 60-67, 2023, <http://dx.doi.org/10.56773/ejer.v2i1.17>
- [2] Gondo, D., Bernardeau-Moreau, D., Campillo, P.: *Student Stress and the Effects of Relaxation: A Study Conducted at the University of Lille in Northern France*, Social Sciences 12(6): 318, 2023, <https://doi.org/10.3390/socsci12060318>
- [3] Nagy, Z.: *Artificial Intelligence and Machine Learning*, Packt Publishing, Birmingham, UK, 2018
- [4] Raschka S., Mirjalili V.: *Machine Learning with PyTorch and Scikit-Learn*, Packt Publishing, Birmingham, UK, 2022
- [5] <https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset>, December 2024.
- [6] Kabita S., Abhaya K.S., Jitendra P., Subhendu K.P.: *Exploratory Data Analysis using Python*, International Journal of Innovative Technology and Exploring Engineering (IJITE) 8(12), 4727-4735, 2019, <https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35911081219.pdf>
- [7] <https://pandas.org>
- [8] <https://scikit-learn.org>
- [9] <https://matplotlib.org>
- [10] <https://seaborn.org>