Muhsin Efe Tarım – Applying ML Methods

**REGRESSION:**

In the third phase of this project, we implemented a linear regression model to predict students’ depression scores based on multiple academic, psychological, and behavioral features. This step builds on the previous exploratory data analysis and hypothesis testing to evaluate how well selected variables can quantitatively explain variations in depression levels.

We selected the following numerical predictors from the dataset:

Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Work/Study Hours, and Financial Stress.

In addition, **we transformed the categorical variable Sleep Duration into an ordinal feature (Sleep\_Ordinal)** to make it suitable for regression analysis. The encoding assigned values from 1 to 4, representing increasing sleep duration from “Less than 5 hours” to “More than 8 hours”.

The model was trained on 80% of the data and tested on the remaining 20%. The evaluation metrics were:

- **Mean Squared Error (MSE)**: 0.1595

- **R-squared (R²)**: 0.3422

A graph with blue squares

AI-generated content may be incorrect.

The regression coefficients revealed the following insights:

• Academic Pressure and Financial Stress were the strongest positive predictors of depression.

• Study Satisfaction showed a mild negative association, suggesting a protective effect.

• Other variables, such as CGPA, had minimal influence.

These findings support the results of the hypothesis testing phase, confirming that economic and academic stressors are more closely linked to depression than academic performance alone.

**CLASSIFICATION:**

In addition to predicting depression scores using regression, we also applied a logistic regression model to classify whether a student is depressed or not. This binary classification task helps to identify at-risk individuals and complements the earlier regression analysis.

Feature Selection and Preprocessing

We selected the following numerical predictors based on previous analyses and exploratory visualizations:

• Academic Pressure

• Work Pressure

• CGPA

• Study Satisfaction

• Job Satisfaction

• Work/Study Hours

• Financial Stress

• Sleep\_Ordinal (transformed from Sleep Duration)

The categorical feature Sleep Duration was converted into an ordinal variable Sleep\_Ordinal to represent increasing hours of sleep (1 = “Less than 5 hours”, …, 4 = “More than 8 hours”). All missing values were dropped.

Model and Evaluation

We used Logistic Regression, trained on 80% of the dataset and tested on the remaining 20%. The following evaluation metrics were used:

• **Accuracy**: 0.78

• **Precision**: 0.75 (Not Depressed), 0.79 (Depressed)

• **Recall**: 0.69 (Not Depressed), 0.84 (Depressed)

• **F1-score**: 0.72 (Not Depressed), 0.81 (Depressed)

• **ROC AUC Score**: 0.84

The confusion matrix showed that the model correctly classified the majority of students, with 2732 true positives (correctly identified as depressed) and 1600 true negatives. While there were some false positives (706) and false negatives (538), the overall balance of precision and recall was strong.

A graph of confusion matrix

AI-generated content may be incorrect.

The ROC curve confirmed the model’s performance, with an AUC of 0.84, indicating a high ability to distinguish between depressed and non-depressed students.

A graph of a logistic regression

AI-generated content may be incorrect.

As a result the classification model performed well, particularly in identifying depressed students (recall = 0.84). These results align with the regression analysis, further reinforcing the impact of academic pressure, financial stress, and sleep habits on mental health. This model could serve as a useful tool in early detection systems for student well-being interventions.