Project Report: Predicting Mood Scores from Digital Habits

1. Problem Description

In this digital age, understanding the relationship between our online habits and mental well-being is crucial. This project aims to address this issue by developing a machine learning model to predict an individual's mood. The target variable for this analysis is the ***mood_score***, a numerical rating from 1 to 10, indicating a person's emotional level. The primary goal was to determine if and how factors like screen time, social media usage, sleep patterns, and stress levels could be used to effectively predict this score.

2. Dataset and Feature Selection

The analysis was conducted using the "*digital_habits_vs_mental_health*" dataset, sourced from Kaggle. The dataset contains self-reported data from numerous individuals, providing a rich source of information for this study.

From the dataset, the following four features were selected as predictors for the model:

- screen_time_hours
- social_media_platforms_used
- sleep_hours
- stress_level

This selection was made to capture distinct and influential aspects of an individual's daily life. The ***sleep_hours*** and ***stress_level*** are well-established psychological factors affecting mood. The ***screen_time_hours*** and ***social_media_platforms_used*** were chosen to represent the quantity and variety of digital engagement. A notable decision was the exclusion of the ***hours_on_TikTok*** feature. This was done to avoid *multicollinearity*, as time spent on a specific platform is inherently a component of total screen time.

3. Modeling Approach

A **Polynomial Regression** model was implemented to capture potentially complex, non-linear relationships between the features and the mood score, The entire workflow was encapsulated in a scikit-learn **Pipeline** to ensure consistency and prevent data leakage. The pipeline consisted of three sequential steps:

1. **StandardScaler**: This step normalized the features, transforming them to have a mean of 0 and a standard deviation of 1.
2. **PolynomialFeatures**: This step generated new features by creating higher-order terms (e.g., x2,x3) and interaction terms (e.g., x1·x2) from the original scaled features.
3. **LinearRegression**: The final step was a standard linear regression model that made predictions based on the set of polynomial features.

To find the optimal parameter of the model, **GridSearchCV** was utilized to perform hyperparameter tuning on the degree of the polynomial. Using 5-fold cross-validation, it systematically tested degrees from 2 to 9 and identified a **degree of 4** as the optimal choice, balancing model fit and complexity.

4. Evaluation

The model's performance was evaluated using the standard regression metrics on both the training and test datasets. The results on the **unseen test data** are the most important indicator of the model's real-world performance.

- **R-squared (R^2)**: The model achieved a test R^2 score of **0.647**. This indicates that our model can explain approximately 64.7% of the variance in the mood scores, which is a reasonably strong result for a complex social science problem.
- **Mean Absolute Error (MAE)**: The test MAE was **0.52**. This means that, on average, the model's predictions were off by about half a point on the 1-to-10 mood scale.

Significantly, the performance metrics on the training set (*R^2 of 0.652, MAE of 0.521*) were nearly identical to the test set scores. This proves that the model generalizes well to new data and avoids overfitting.

5. Reflection

This project provided several key insights and learning opportunities. The primary challenge was in the initial feature selection phase, requiring careful consideration to avoid **multicollinearity** and select a set of logically independent predictors.

A significant lesson from this project was the practical application of a structured machine learning workflow. Using a **Pipeline** to link preprocessing and modeling, combined with **GridSearchCV** for automated tuning, created a robust and reproducible process. It reinforced that a model's success depends as much on proper data handling and tuning.

Furthermore, the model is limited by the available data. Incorporating additional features, such as the nature of online content consumed, physical activity levels, or dietary habits, would likely provide a more holistic view and yield an even more powerful predictive model.