CSE271: Data Science Methodology
Instructor: Fatma ElShahaby

Social Media and Smartphone Addiction Analysis

Abstract

This project explores the relationship between smartphone usage patterns and indicators of addiction and stress among users. Using behavioral datasets focused on mobile app engagement, we applied both regression and classification models to analyze how features such as screen time, notifications, and night usage influence user well-being. For classification, we developed models to predict addiction status and high stress risk, evaluating their performance using precision, recall, and ROC analysis. For regression, we modeled users' stress level based on multiple behavioral features using multiple regression models to improve generalization. Data visualization techniques, including boxplots, heatmaps, and scatter plots, were employed to uncover trends and correlations in the data. An interactive dashboard was also created to allow dynamic exploration of key insights. The results suggest strong behavioral patterns associated with both addiction and stress risk, offering valuable insights into the impact of digital habits on mental well-being. Our findings highlight the potential of data science in behavioral health monitoring and digital well-being assessment.

Introduction

Background and Context

The rapid growth of smartphones and social media has significantly reshaped how individuals interact, communicate, and consume information. While these technologies have enhanced convenience and connectivity, growing evidence suggests a concerning rise in problematic usage patterns. Excessive screen time, frequent app switching, and constant notifications have been associated with increased stress, reduced attention span, and even addiction-like behaviors. Understanding these patterns is essential to addressing the mental and emotional health implications of smartphone overuse.

Literature Review

Recent studies have explored the links between smartphone addiction and various psychological outcomes, such as anxiety, depression, and sleep disturbances. Many have relied on survey-based methods to measure digital dependence, while fewer have leveraged behavioral datasets with concrete usage metrics. This project builds on existing research by

applying data science techniques to analyze large-scale, app-based behavior data, offering a more objective lens into user habits.

Project Objectives

This project aims to analyze smartphone and social media usage to uncover behavioral indicators of stress and addiction. Our goals are to:

- Explore and visualize usage trends using interactive and static tools.
- Predict addiction and stress risk using classification models.
- Model stress levels using regression analysis.
- Evaluate the influence of key features (e.g., daily screen time, notifications, night usage, age) on user well-being.
- Present insights through a dynamic dashboard and a comprehensive technical report.

By combining machine learning, statistical analysis, and visualization, this project contributes to a deeper understanding of digital wellness and the behavioral markers of smartphone overuse.

Materials and Methods

Data Description

The dataset used in this project, mobile_usage_analysis.csv, contains behavioral and demographic records from 13,589 smartphone users. It includes variables such as screen time, number of notifications, night usage hours, age, gaming time, and number of installed apps. The datasets are numeric in nature and suitable for both regression and classification tasks. The Target Variables are defined:

- **Addiction_status** (binary): Indicating whether a user is addicted to their mobile phone or not addicted.
- **Stress_risk** (binary): Indicating if a user is subject to high stress risks due to overusing their mobile phones.
- **Stress_level** (continuous): Indicating the level of stress users are subject to, due to using their mobile phones.

Data Preprocessing

- All features were checked for missing values. None were found.
- The addiction_status field was converted into binary values instead of string values.
- The stress_risk was calculated into binary values instead of continuous values.
- Explanatory variables were scaled using StandardScaler() to prepare for models sensitive to feature magnitude.
- Data was split into 80% Training and 20% testing ratios.

Methodology

Visualization Techniques

Exploratory Data Analysis was conducted using static and interactive visualizations:

- **Boxplots** were used to examine features related to binary target variables.
- **Heatmaps** revealed correlations between numeric features and target variables.
- **Bar Charts** were used to measure feature importance in regression models.
- **Scatter Plots** were used to model relationships between predicted values and actual test values across the different models. Additionally, scatter plots were used to plot residual plots.
- **Pairplot** illustrating the pairwise relationships between key variables, with data points color-coded by target variables to reveal potential clustering and trends across different user behaviors.

Modeling Approaches

Classification

Two models were trained to predict:

- **Addiction_status** (Addicted vs Not addicted)
- **Stress_risk** (High vs Low Stress Risk)

Models used:

- **Logistic Regression**: a parametric, model-based, supervised learning algorithm. It was selected for its simplicity and interpretability in binary classification.
- **K-Nearest Neighbours**: a non-parametric, instance-based, supervised learning algorithm. It was selected for its flexibility and ability to catch non-linear patterns.

They were chosen together to provide variety and contrast in modeling approaches.

Regression

Four models were trained to predict the **stress levels** based on their **daily screen time** and their **night usage**

Models Used:

- **Multiple Linear Regression:** parametric model, selected based on exploratory analysis using correlation heatmaps that showed the linearity between the target variable “stress_level” and the selected features.
- **Regularised Linear Regression (Lasso):** parametric model, selected to enhance the performance of the linear model by eliminating unimportant features and penalizing large coefficients to avoid overfitting.
- **Polynomial Regression:** parametric model, selected to capture slightly more complex relationships, if any, but still assumes a predictable mathematical pattern (polynomial).
- **Random Forest Regression:** non-parametric model, selected for a more complex analysis using step-like approximations.

Model Evaluation

Each model was assessed using appropriate evaluation metrics:

Classification

Accuracy, Precision, Recall, F1-Score, and ROC Curves. Confusion matrices were also presented.

Regression

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), r^2 (Coefficient of Determination), residual plots, and Prediction vs Actual value plots.

All models were trained on 80% of the data and tested on the remaining 20% to assess generalization performance.

Results

Key Findings

Significant patterns and trends

Exploratory analysis revealed that higher daily screen time, night usage, and notifications are strongly associated with both addiction and higher stress levels. Younger users also showed stronger tendencies to higher stress risks.

Classification Models Analysis

The Logistic regression model yielded the following confusion matrix:

		Predicted	
		Not addicted	**Addicted**
Actual	**Not Addicted**	1306	30
	Addicted	24	1358

Providing the following metrics:

```
Accuracy:    0.9801324503311258
Precision:   0.978386167146974
Recall:      0.9826338639652678
F1 Score:    0.9805054151624548
```

The K-Nearest Neighbours model yielded the following confusion matrix:

		Predicted	
		Not addicted	**Addicted**
Actual	**Not Addicted**	1785	52
	Addicted	28	853

Providing the following metrics:

```
Accuracy:    0.9705665930831494
Precision:   0.9425414364640884
Recall:      0.9682179341657208
F1 Score:    0.9552071668533034
```

Regression Models Analysis

By inspecting feature importance using bar charts, measuring coefficients, and looking at feature importance, it was found that the most important feature is `daily_screen_time`.

```
Linear Regression Coefficients:
night_usage          -0.259334
daily_screen_time    17.549846

Lasso Regression Coefficients:
night_usage          -0.085501
daily_screen_time    17.375891

Polynomial Regression Coefficients:
night_usage                       -0.265930
daily_screen_time night_usage     -0.205116
daily_screen_time^2               -0.052320
1                                  0.000000
night_usage^2                      0.054161
daily_screen_time                 17.558414

Random Forest Feature Importances:
night_usage          0.102825
daily_screen_time    0.897175


Linear Regression R²: 0.7573590356729519
Linear Regression MSE: 98.660183719092
Linear Regression RMSE: 9.93278328159293


Lasso Regression R²: 0.757460789708498
Lasso Regression MSE: 98.61880953535113
Lasso Regression RMSE: 9.9307003547258
```

```
Polynomial Regression Training R²: 0.7577919580370938
Polynomial Regression Training MSE: 97.24666813567437
Polynomial Regression Training RMSE: 9.86137252798384


Polynomial Regression Testing R²: 0.7573384937058408
Polynomial Regression Testing MSE: 98.66853628336222
Polynomial Regression Testing RMSE: 9.933203727064205


Random Forest R²: 0.661072397529568
Random Forest Regression MSE: 137.81127032669264
Random Forest Regression RMSE: 11.739304507793152
```

Visualizations

Classification

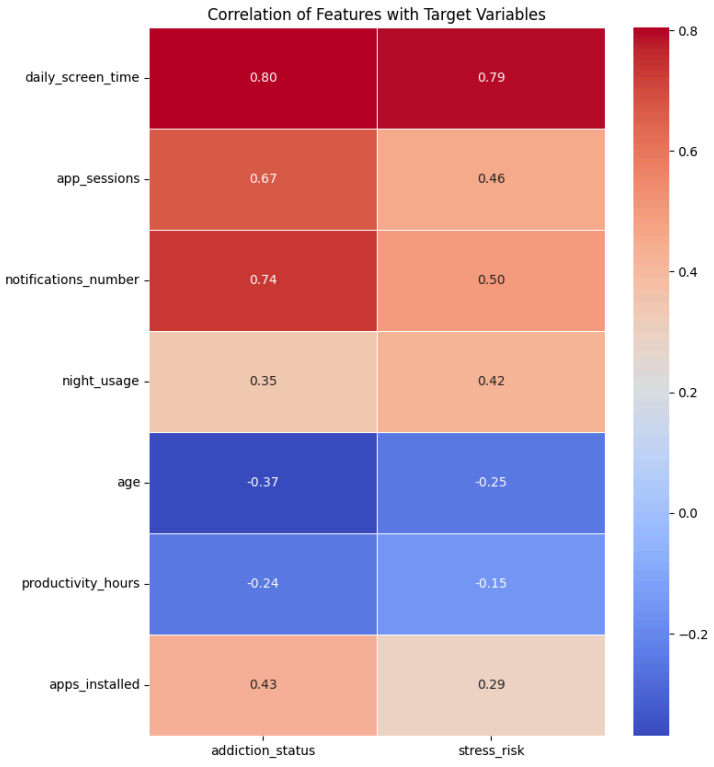

Figure 1.1

Correlation Heatmap of Features to Target Variables

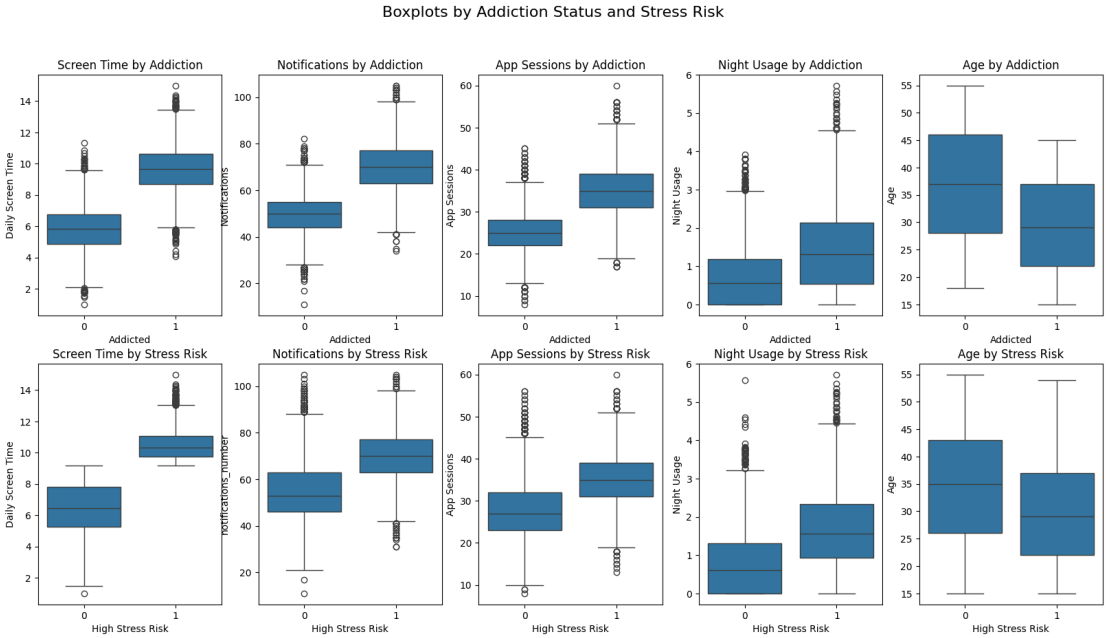

Figure 1.2
Boxplots illustrating relations between key Features and Target Variables

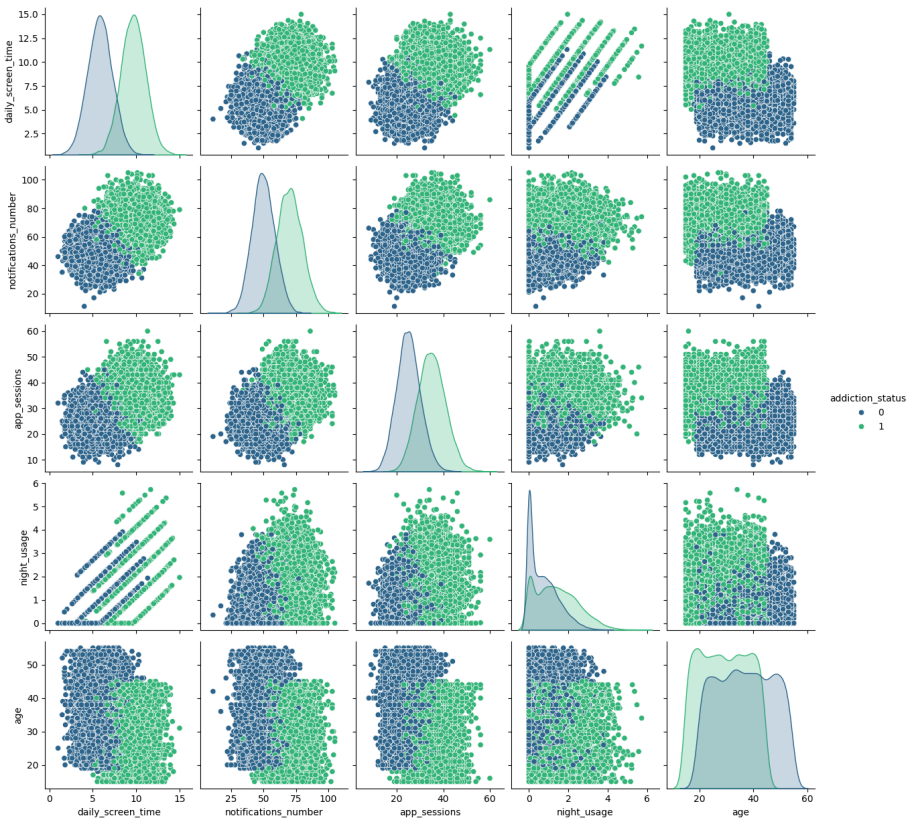

Figure 1.3
Pairplot illustrating the pairwise relationships between key features, with data points color-coded by addiction status to reveal potential clustering and trends across different user behaviors

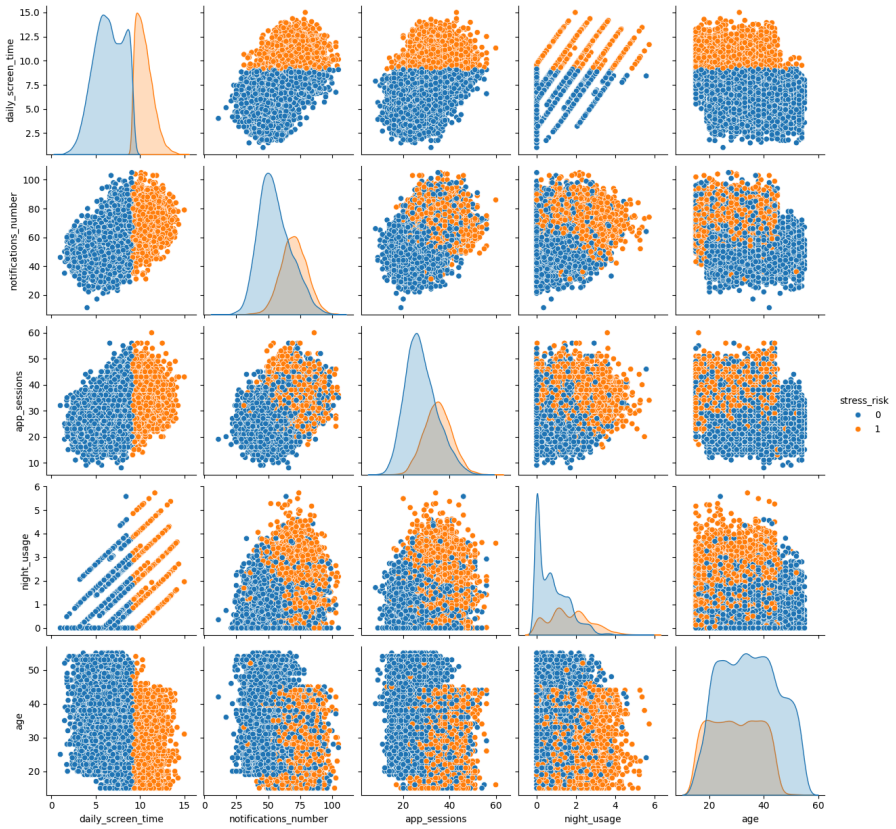

Figure 1.4

Pairplot illustrating the pairwise relationships between key features, with data points color-coded by stress risk to reveal potential clustering and trends across different user behaviors

---

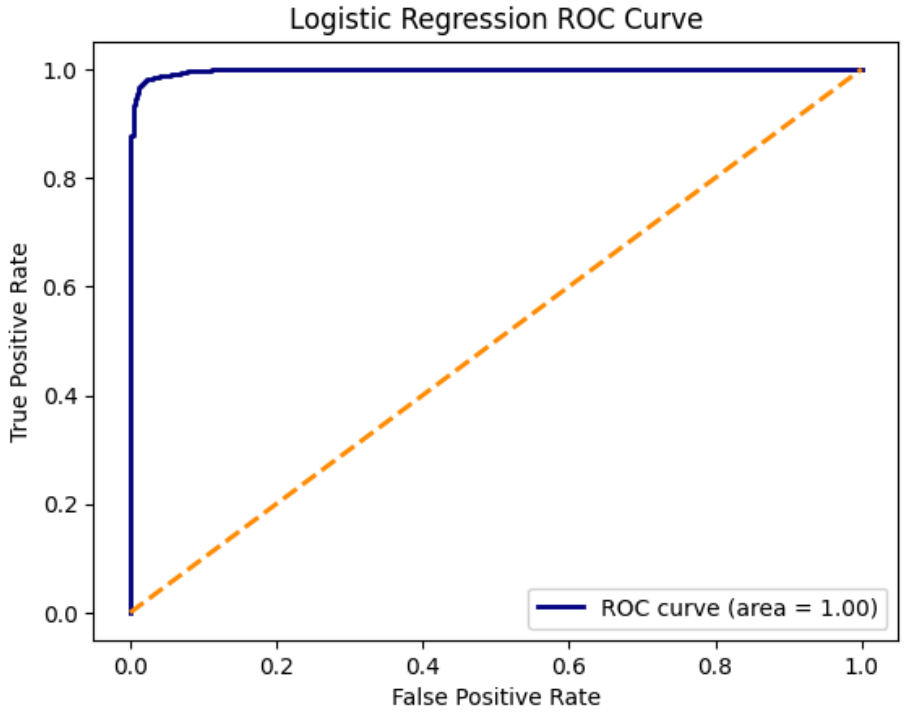

Figure 1.5

ROC Curve of the Logistic Regression Model

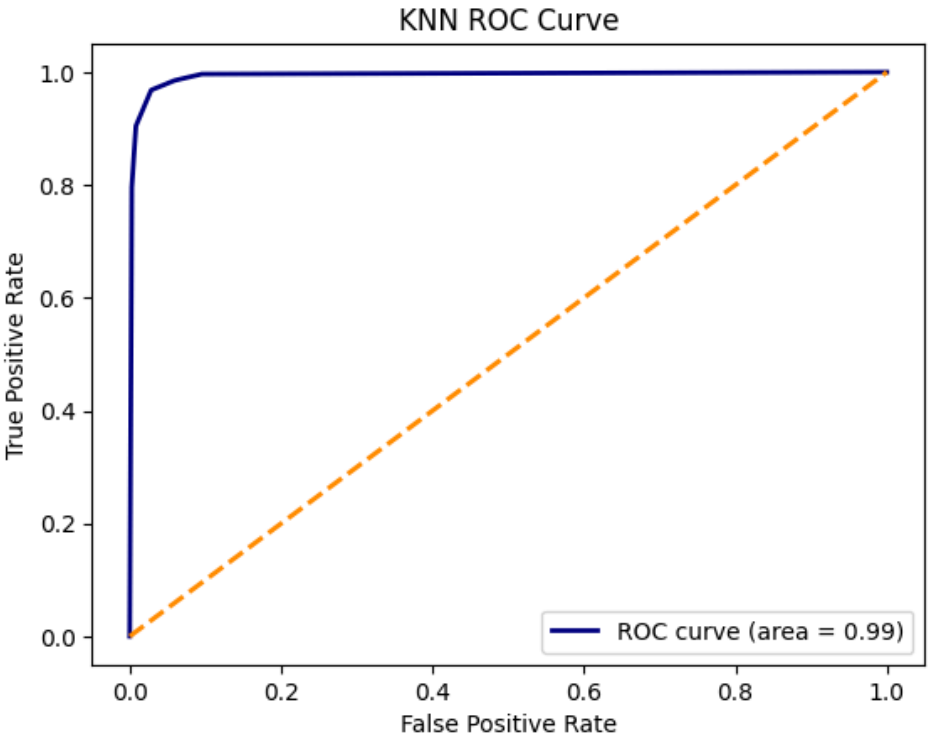

Figure 1.6
ROC Curve of the KNN Model

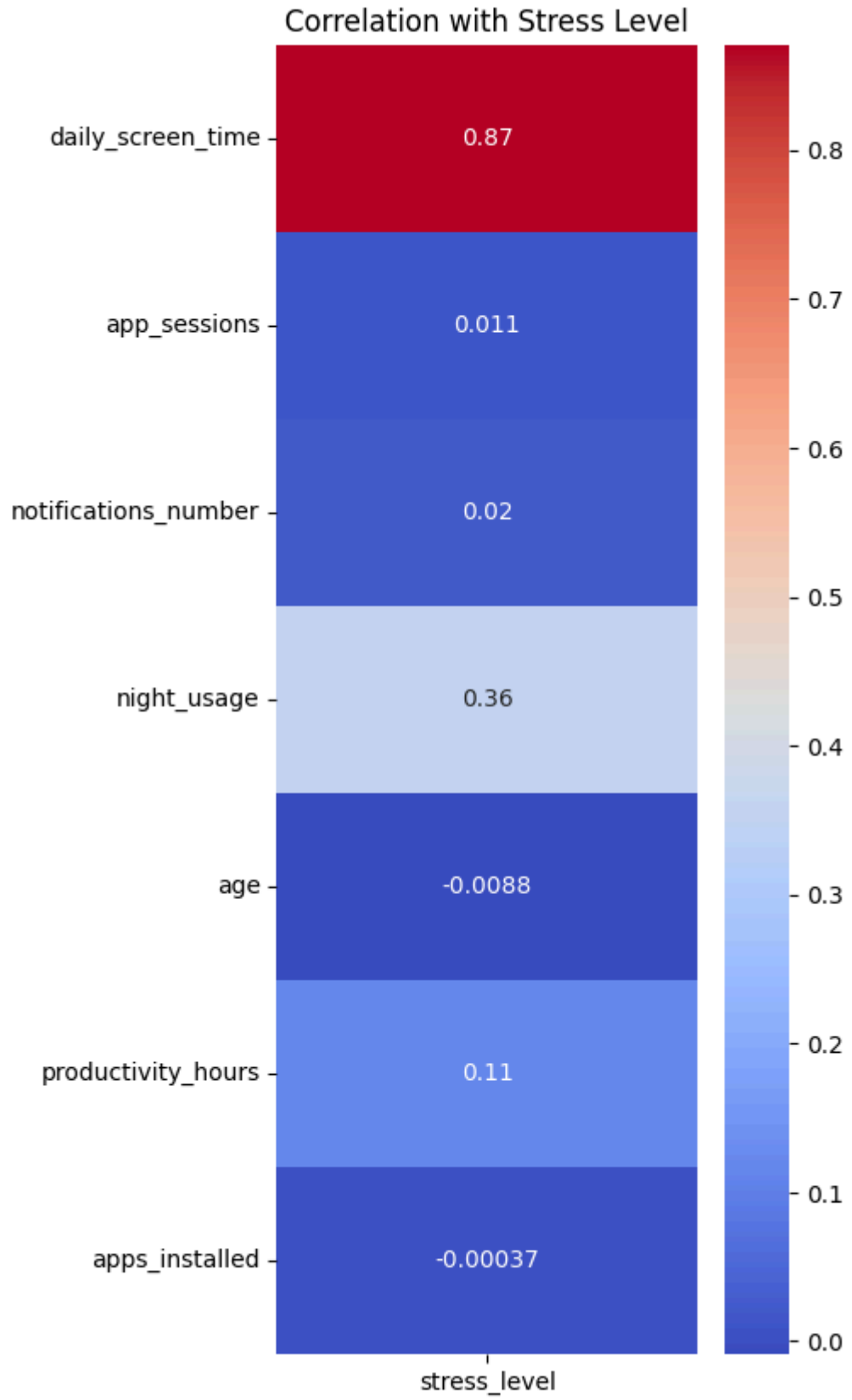

Figure 2.1
Heatmap showing correlations between stress level and different variables

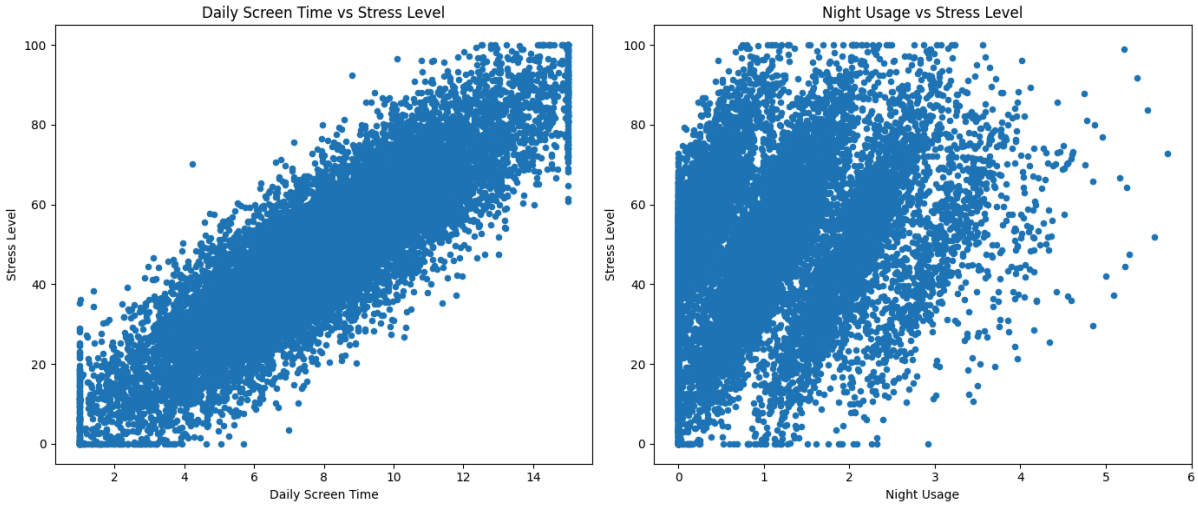

Figure 2.2

Scatter plot showing the relationship between daily screen time and night usage with stress level

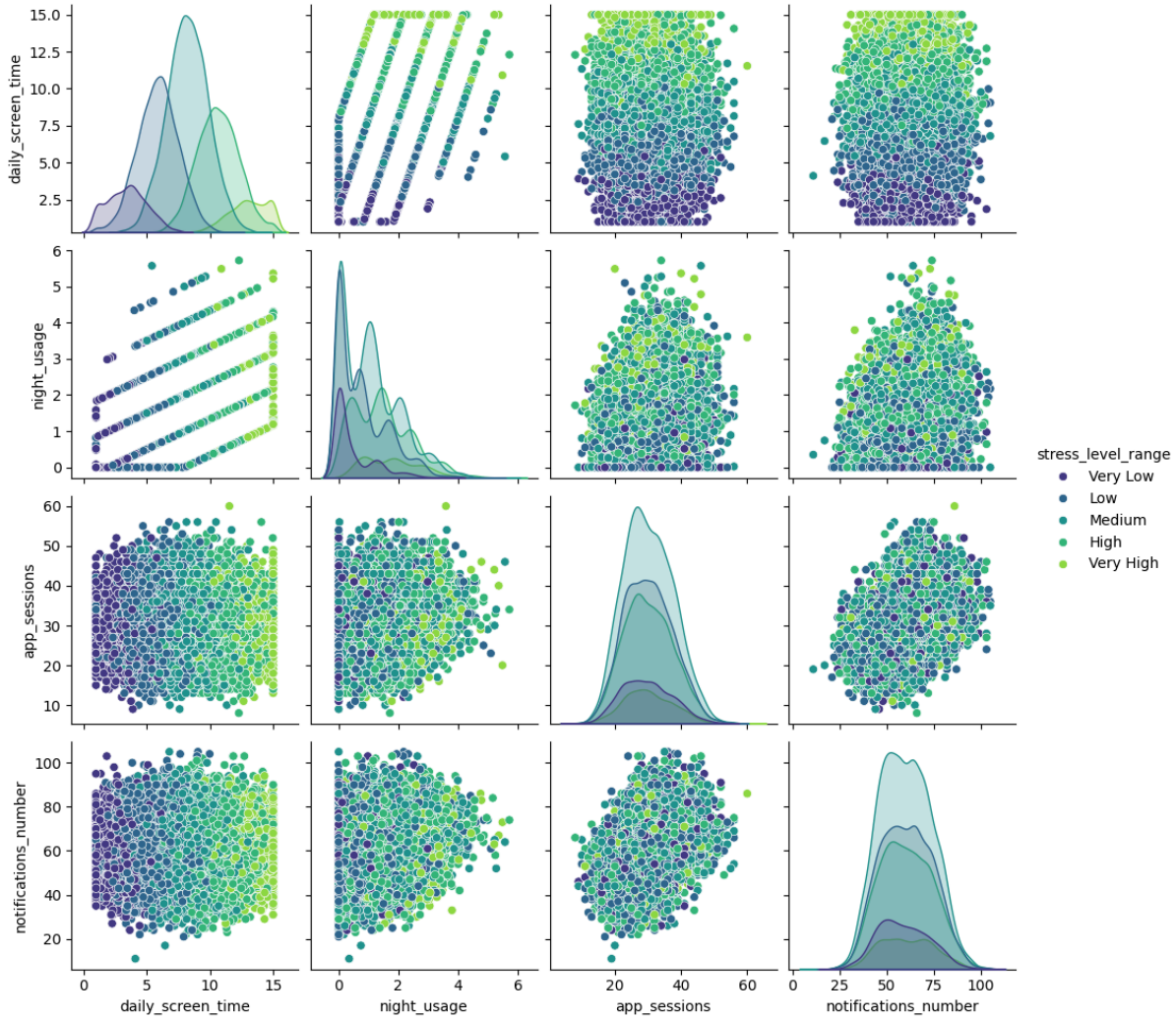

Figure 2.3

Pairplot illustrating the pairwise relationships between key variables, with data points color-coded by stress level ranges to reveal potential clustering and trends across different user behaviors

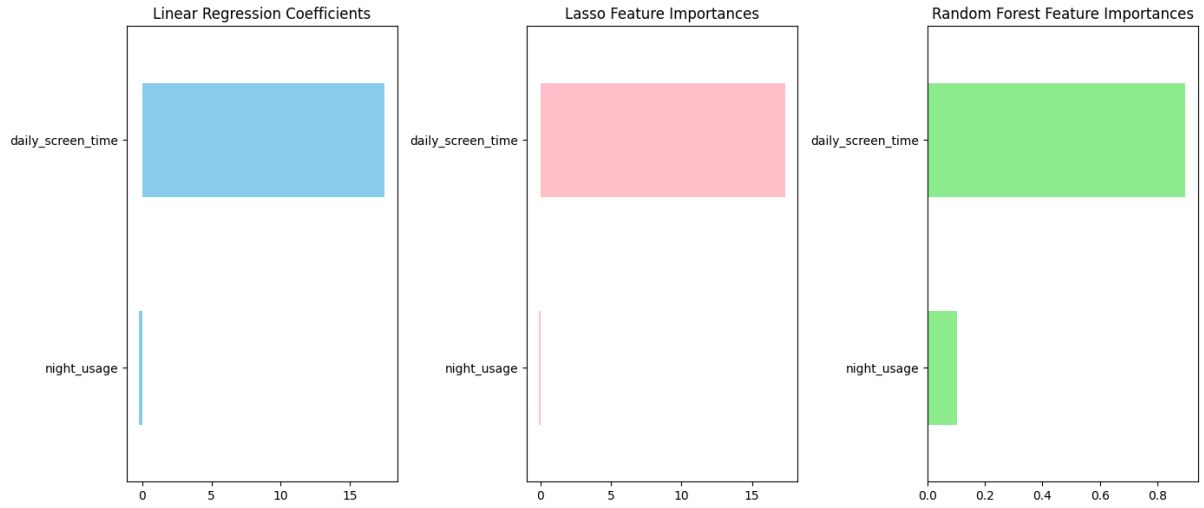

Figure 2.4

Bar Charts showing feature importance for each regression model

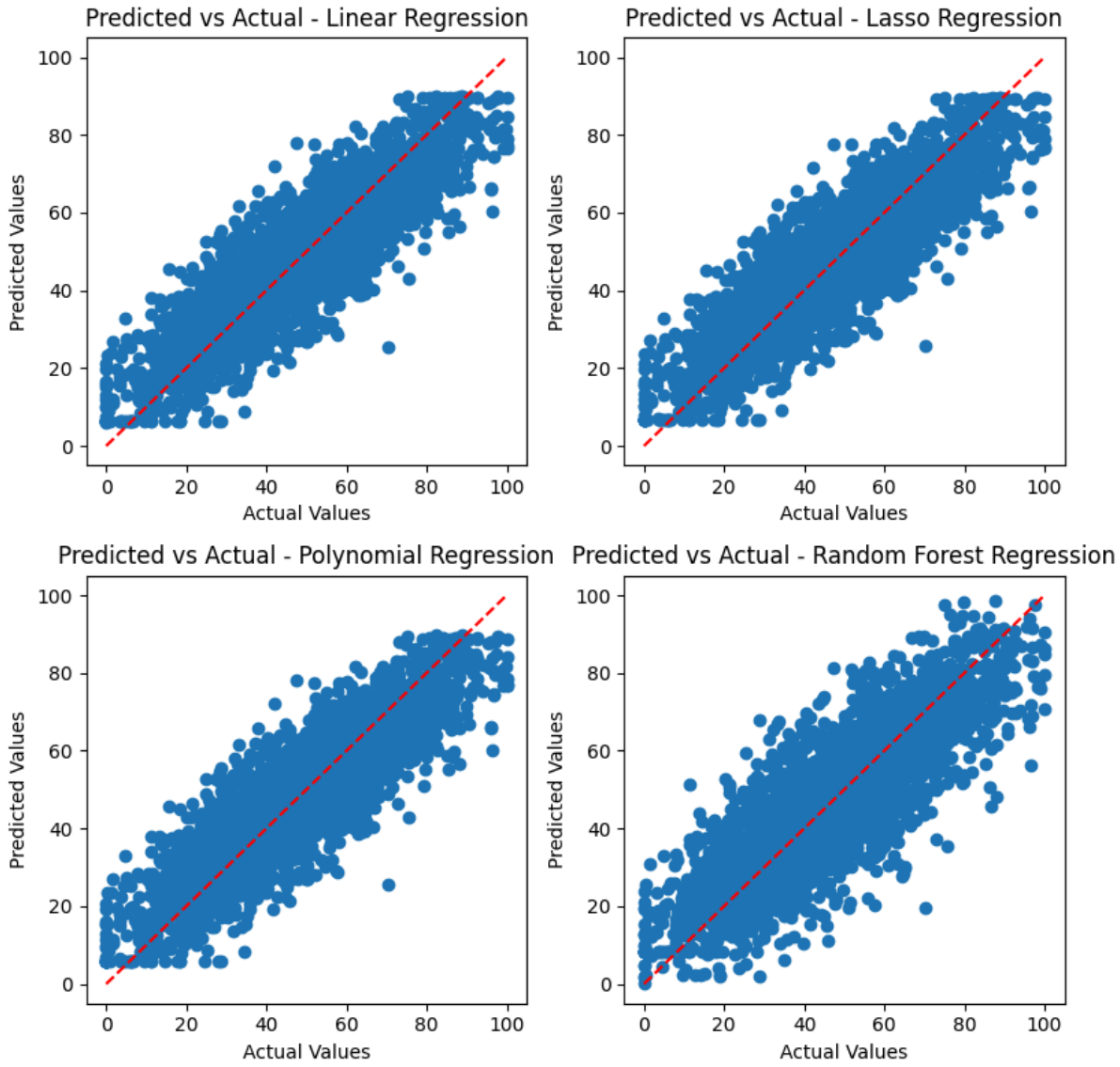

Figure 2.5

Scatter plot showing the relationship between predicted values versus Actual Values for each regression model

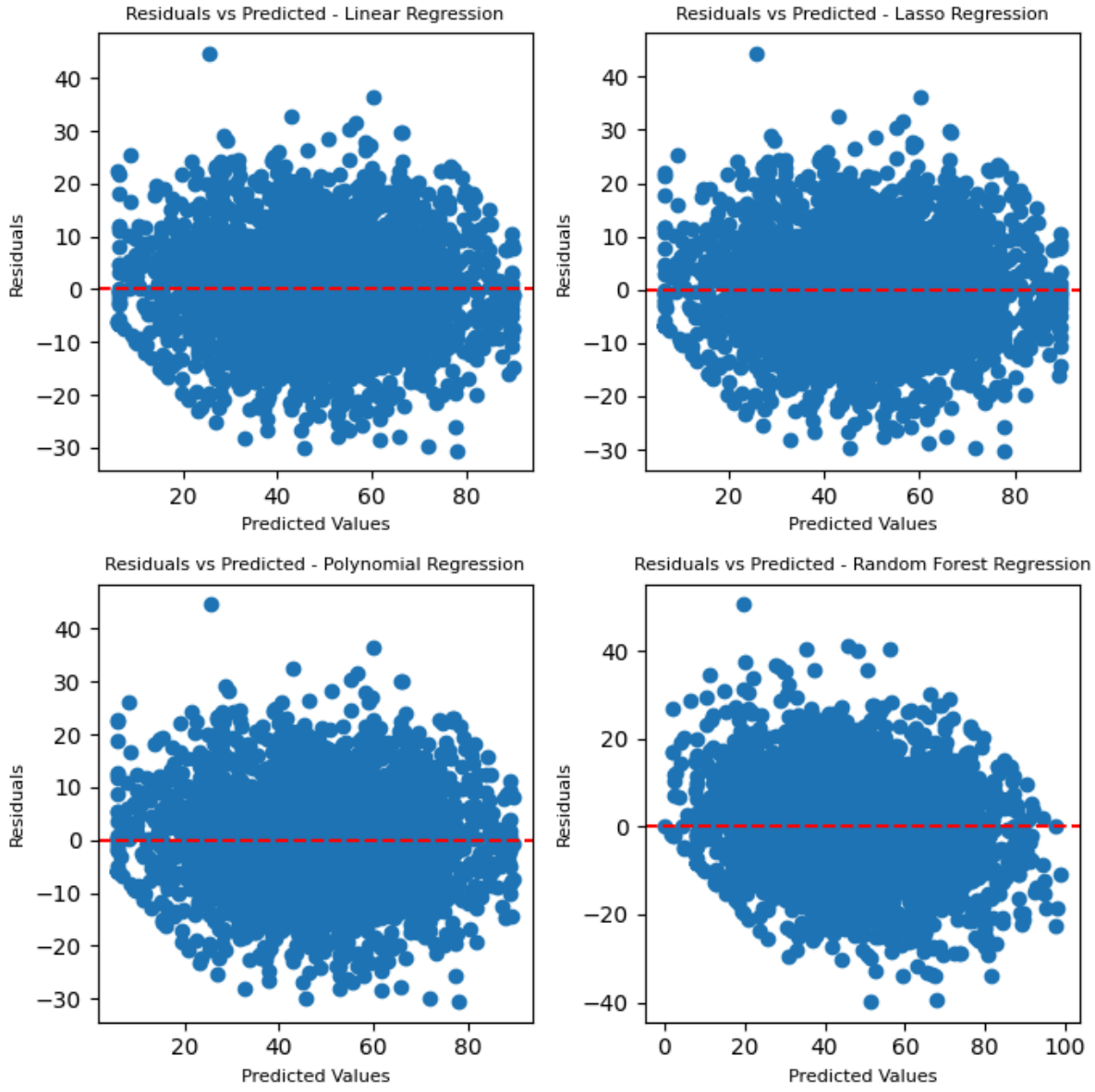

Figure 2.6
Residual plot to evaluate the difference between model predictions and actual values in the testing dataset

Model Performance

Classification

Both classification models achieved high accuracy, with **Logistic Regression** performing slightly better overall. It reached an **accuracy of 98%**, with **precision (97.8%)** and **recall (98.3%)**, indicating it was highly effective at identifying addicted users with minimal misclassification.

The **K-Nearest Neighbors (KNN)** model, used for predicting stress risk, also performed well, achieving **97% accuracy**. It showed strong recall (96.8%) and precision (94.3%), meaning it reliably detected high-stress users with few false positives.

Despite using different approaches—logistic regression as a linear model and KNN as a distance-based method—both models aligned closely with the underlying structure of the data and confirmed the strength of the selected features.

Regression

Multiple Linear Regression, Lasso Regression, and Polynomial Regression Models

All three models show almost identical performance, with $r^2 \approx 0.757$ and RMSE ≈ 9.93. This suggests that the relationship between the features and the target variable is predominantly linear (as shown in the pairplot).

Residual plots confirm this, as they show random scatter around zero, indicating no obvious patterns or systematic bias.
Lasso Regression performs similarly to Linear Regression, implying that regularization had little effect. This is because daily screen time has a significantly stronger linear relationship to stress level than night usage.

Additionally, adding polynomial features did **not significantly improve performance**, which further confirms that the **underlying relationship is linear**.

The **training and testing r^2 are consistent for the polynomial model**, indicating no overfitting.

Random Forest Regression model

Despite producing a visually similar predicted vs. actual scatter plot, the **r^2 is lower (0.661),** and **RMSE is higher (11.74)**.

The **residuals are unevenly distributed** and show a banding pattern typical of tree-based models approximating continuous linear relationships.

This means Random Forest **fails to capture the smooth linear trend** and instead **overfits local variations**, reducing overall performance.

Conclusions

Summary of Findings

From the analysis, smartphone usage patterns were found to be strong predictors of both stress levels and addiction risk. Logistic Regression achieved 98% accuracy in identifying addicted users, while KNN reached 97% accuracy for stress risk classification. In regression, all linear models performed similarly with r^2 around 0.76 and RMSE 9.93, while the Random Forest model had a slightly lower r^2 of around 0.66 and a slightly higher RMSE of around 11.74, indicating a primarily linear relationship between behavioural features and stress. These findings highlight the potential of usage data as a reliable indicator of user well-being.

Recommendations

- Prioritize interventions targeting night-time phone usage, high screen time, and frequent app sessions, as these factors showed the strongest relationships with indicators of addiction.

- For digital well-being apps or awareness programs, focus on reducing notification frequency and promoting screen breaks, as both are correlated with increased screen time.

- Encourage age-specific approaches, especially for younger users who may have more vulnerable usage patterns.

Limitations

- **Feature Independence Assumption**: Logistic Regression assumes linear relationships and limited feature interaction, which may not hold in real behavioral data. Similarly, KNN is sensitive to feature scaling and density, which could affect performance in more diverse datasets.

- **Lack of Temporal Data**: The dataset captures usage behavior at a single point in time. Trends over time (e.g., increasing stress or addiction) could not be analyzed.

- **No External Validation**: The models were evaluated only on data from the same source, with no external dataset used to test generalizability.

- **No Cross Validation Used:** The results are based on a single test/train split; cross-validation could provide more robust estimates of model performance.

Future Work

- Perform cross-validation to ensure that model performance is consistent across different data splits.

- Other relationships can be explored, for instance, how age is related to variables such as screen time, night usage, and stress levels.
- Also, a potential avenue could be exploring how the number of notifications per day and the number of app sessions affect productivity hours.
- Testing models on real-world or unseen datasets to assess their generalizability.

Acknowledgements

Team Contributions:

Rokaya Ramy was responsible for the Regression analysis, including building and evaluating the Linear, Polynomial, Lasso, and Random Forest models to predict stress levels.

Jana Sherif handled the Classification models, including Logistic Regression and K-Neares Neighbours, and evaluated their performance in predicting users' addiction status and risks of stress.

External Support

This project relied on several external resources and tools:

- **Libraries**: Python libraries such as `pandas`, `numpy`, `scikit-learn`, `seaborn`, `matplotlib`, and `streamlit` were essential for data analysis, modeling, visualization, and dashboard development.
- **Datasets:** The project used the publicly available datasets `mobile_addiction.csv` and `mobile_usage_analysis.csv`.
- **Mentors:** We gratefully acknowledge **Dr. Fatma** for her guidance and support throughout the course. All her efforts and assistance are highly appreciated.

References

GeeksforGeeks. (2025, March 20). *Mean squared error in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/python-mean-squared-error/

Pande, K. (2024, August 23). *How to read a heatmap: A Beginner's guide*. Website. https://vwo.com/website-heatmap/how-to-read-heatmap/#:~:text=Reading%20a%20heat%20map%20is,and%20green%20signify%20low%20values

GeeksforGeeks. (2025b, April 7). *Random forest regression in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-regression-in-python/