Title of the Project:

Dopamine Overload Prediction

Abstract:

In an increasingly digitalized world, the pervasive integration of technology into daily life, while offering substantial benefits, also poses significant challenges to mental well-being. One critical yet often overlooked consequence of continuous digital stimulation is the potential for dopamine overload, wherein frequent notifications, instant gratification mechanisms, and prolonged screen engagement may dysregulate the brain's reward system. This dysregulation has been associated with adverse psychological outcomes, including heightened anxiety, stress, depressive symptoms, and diminished attention and focus.

This project addresses the growing challenge of identifying individuals at risk of digital dependence and dopamine-related reward dysregulation. Traditional mental health assessment approaches often fail to capture the cumulative and subtle effects of digital lifestyle behaviors, limiting opportunities for early intervention and prevention. As a result, individuals may remain vulnerable to the long-term psychological impacts of excessive digital engagement.

To bridge this gap, the development of a predictive framework for identifying high-risk digital lifestyle patterns indicative of potential dopamine overload is proposed. Leveraging a comprehensive dataset encompassing digital behavior metrics, mental health indicators, and lifestyle factors, the project aims to identify key risk determinants and construct predictive models capable of classifying individuals according to risk level. The findings are expected to contribute to a deeper understanding of the relationship between digital behavior and mental health, while offering practical implications for personalized interventions and public health strategies designed to promote healthier and more balanced digital engagement.

Introduction and Motivation

This project focuses on identifying and predicting instances of digital dependence or 'dopamine overload' based on an individual's digital lifestyle habits and mental wellness indicators.

In our increasingly digital world, individuals are constantly exposed to stimuli through devices and social media, which can lead to overstimulation and digital dependence, commonly referred to as 'dopamine overload.' This phenomenon can significantly impact mental health, productivity, and overall well-being. Predicting the risk of such overload is crucial for early intervention and for promoting healthier digital habits. By understanding the factors contributing to this state, we can develop strategies to mitigate its negative effects.

To achieve this, the project utilizes the Digital Lifestyle Benchmark dataset, sourced from Kaggle. This dataset contains a comprehensive collection of information on device usage, social media engagement, and various mental health scores, providing a rich foundation for developing predictive models.

Objectives

The primary goals of this project are twofold: firstly, to develop and evaluate classification models capable of predicting the `high_risk_flag`, indicating individuals at high risk of dopamine overload. Secondly, to construct and assess regression models for accurately predicting the `digital_dependence_score`, quantifying the level of digital reliance. These objectives are directly derived from the machine learning models that have been developed and evaluated within this project, serving to identify and quantify critical aspects of digital well-being.

Other objectives include:

- **Identify Key Risk Factors:** Analyze the provided dataset to uncover significant demographic, behavioral, and psychological factors contributing to digital dependence and dopamine overload.
- **Evaluate Model Performance:** Rigorously assess the performance of all developed models using appropriate metrics (e.g., Accuracy, Precision, Recall, F1-score, ROC-AUC for classification; MSE, R2 for regression).
- **Analyze Feature Importance:** Utilize techniques like SHAP values to understand which features are most influential in predicting both dopamine overload risk and digital dependence.

Methodology

Data Collection

The dataset utilized for this analysis is the **Digital Lifestyle Benchmark** dataset, sourced from HF/Kaggle. It provides comprehensive data on various aspects of digital habits, mental well-being and productivity. The dataset was loaded into a pandas DataFrame named `df` using the `pd.read_csv()` function.

Upon initial inspection, the dataset contains 3500 entries and 24 columns, with no missing values detected across any of the features, as confirmed by `df.isnull().sum()`. The columns comprise a mix of 11 float64, 7 int64, and 6 object (categorical) data types.

Further exploration of the dataset's basic characteristics revealed the following:

- **Numerical Features**: Descriptive statistics for numerical columns (`df.describe()`) showed a wide range of values, indicating variability in metrics such as `age`, `device_hours_per_day`, `phone_unlocks`, `notifications_per_day`, and various mental health scores (`anxiety_score`, `depression_score`, `stress_level`). For instance, `device_hours_per_day` ranges from approximately 0 to 16 hours, while `phone_unlocks` can be as high as 300+ times per day. The `digital_dependence_score` also varies significantly.
- **Categorical Features**: The dataset includes categorical columns such as `gender`, `region`, `income_level`, `education_level`, `daily_role`, and `device_type`. These features have a manageable number of unique categories, for example, `gender` has 2 unique values (Female, Male), `region` has 6 unique values (Europe, Asia, North America, Africa, South America, Middle East), and `income_level` has 4 unique values (Low, Lower-Mid, Upper-Mid, High). The distribution within these categories provides insights into the demographic and behavioral composition of the individuals in the dataset.

This initial overview confirms the dataset's structure and completeness, preparing it for subsequent preprocessing and analysis.

Data Preprocessing and Feature Engineering

Data Cleaning

Initial checks revealed no duplicate rows in the dataset, thus `df.drop_duplicates()` did not alter the DataFrame size. Similarly, there were no missing values across any of the columns, negating the need for imputation techniques. This indicates a clean dataset from the outset, allowing us to proceed directly to feature engineering and transformation.

Feature Engineering

Four new features were created to capture more nuanced aspects of digital behavior and mental well-being:

- **`screen_time_per_unlock`**: Calculated as `(device_hours_per_day * 60) / (df_clean['phone_unlocks'] + 1e-6)`. This feature aims to quantify the average duration between phone unlocks, potentially indicating usage patterns (e.g., long continuous sessions vs. frequent short checks). A small constant (`1e-6`) was added to the denominator to prevent division by zero.

- **`notifications_per_hour`**: Calculated as `notifications_per_day / (df_clean['device_hours_per_day'] + 1e-6)`. This metric provides insight into the intensity of digital interruptions relative to device usage time, which could be a proxy for digital overload. A small constant (`1e-6`) was added to the denominator to prevent division by zero.
- **`mental_health_load`**: Sum of `anxiety_score`, `depression_score`, and `stress_level`. This composite score provides a holistic view of an individual's overall mental health burden.
- **`mood_balance`**: Calculated as `happiness_score - stress_level`. This feature attempts to capture the equilibrium between positive and negative emotional states, where a lower or negative value might indicate an imbalance.

Feature Categorization

Features were categorized into two main groups for distinct preprocessing:

- **`numeric_features`**: All numerical columns (int64, float64), excluding 'id' and the target variable 'high_risk_flag'. This includes original numerical features and the newly engineered ones.
- **`categorical_features`**: All object-type columns, representing nominal or ordinal data.

Data Splitting

The dataset was split into training and testing sets using `train_test_split` with a `test_size` of 25%. Critically, `stratify=y` was used to ensure that both the training and testing sets maintained the same proportion of the target variable (`high_risk_flag`) as the original dataset. This is crucial for imbalanced datasets to ensure representative samples in both sets.

Preprocessing Pipeline Setup

A `ColumnTransformer` was employed to build a robust preprocessing pipeline:

- **Numeric Features**: A `StandardScaler` was applied to numeric features to standardize their ranges, which is essential for many machine learning algorithms that are sensitive to feature scales (e.g., Logistic Regression, SVM, kNN).
- **Categorical Features**: A `OneHotEncoder` was used for categorical features, converting them into a binary matrix representation. `handle_unknown='ignore'` was set to manage unseen categories in the test set gracefully, and `sparse_output=False` ensured a dense array output.

This `preprocess` pipeline was then fitted on `X_train` and used to transform both `X_train` and `X_test`, resulting in `X_train_t` and `X_test_t`.

Addressing Class Imbalance with SMOTE

The target variable (`high_risk_flag`) exhibited a significant class imbalance (approximately 80% non-risk, 20% high-risk). To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data only (`X_train_t`, `y_train`). SMOTE generates synthetic samples for the minority class, effectively balancing the class distribution and preventing models from being biased towards the majority class. This resulted in a balanced `X_train_smote` and `y_train_smote`.

Model Training

Classification Models (Predicting high_risk_flag)

To handle the class imbalance observed in the `high_risk_flag` target variable, all classification models were trained on the SMOTE-resampled training data (`X_train_smote`, `y_train_smote`).

- **Logistic Regression**: A linear model used for binary classification. It estimates the probability of an instance belonging to a particular class by fitting data to a logistic function. Its purpose in this project is to serve as a baseline, interpretable classification model to predict whether an individual is at `high_risk_flag` (dopamine overload).
- **K-Nearest Neighbors (kNN)**: A non-parametric, instance-based learning algorithm that classifies a data point based on how its neighbors are classified. It looks at the K closest data points and assigns the class that is most common among them. In this project, kNN is used to explore a non-linear, distance-based approach to classify `high_risk_flag`.
- **Random Forest**: An ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It's known for its high accuracy and robustness. Here, it is employed to capture complex relationships within the data for predicting `high_risk_flag`.
- **XGBoost (Extreme Gradient Boosting)**: An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost is used in this project as a powerful and highly performant model to predict `high_risk_flag`, often achieving state-of-the-art results.
- **Support Vector Machine (SVM)**: A supervised learning model used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space, maximizing the margin between the classes. SVM is utilized to find an optimal decision boundary for classifying `high_risk_flag`.

Regression Model (Predicting digital_dependence_score)

- **RidgeCV (Linear Regression)**: A type of linear regression that includes an L2 regularization term. This regularization helps to prevent overfitting by penalizing large coefficients, making it suitable for datasets with multicollinearity. The `CV` in `RidgeCV` indicates that it performs cross-validation to select the best regularization strength (alpha). In this project, RidgeCV is used to predict the continuous `digital_dependence_score`, aiming for a robust and generalized linear prediction.

Results and Discussion

Classification Models

Performance Metrics Overview

The `comparison_df` DataFrame below summarizes the performance metrics (Accuracy, Precision, Recall, F1 Score) for Logistic Regression, K-Nearest Neighbors (kNN), Random Forest, XGBoost Classifier, and Support Vector Machine (SVM) Classifier.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.699	0.363	0.653	0.467
kNN Classifier	0.565	0.275	0.710	0.396
Random Forest	0.870	0.772	0.500	0.607
XGBoost Classifier	0.872	0.767	0.523	0.622
SVM Classifier	0.800	0.503	0.557	0.528

Overall Performance Trends

From the comparison table and the performance visualization, we can observe distinct trends:

- **K-Nearest Neighbors (kNN)** performed the worst across most metrics, notably having the lowest Accuracy (0.565) and F1 Score (0.396), despite a relatively high Recall (0.710). Its low Precision (0.275) indicates a high number of false positives.
- **Logistic Regression** showed slightly better performance than kNN, with an Accuracy of 0.699 and an F1 Score of 0.467. It also achieved a decent Recall (0.653) but suffered from low Precision (0.363).
- **Support Vector Machine (SVM)** performed moderately well, with an Accuracy of 0.800 and an F1 Score of 0.528. It provided a better balance between Precision (0.503) and Recall (0.557) compared to Logistic Regression and kNN.
- **Random Forest** and **XGBoost Classifier** emerged as the top-performing models. Both achieved high Accuracy (around 0.870), with XGBoost Classifier slightly edging out

Random Forest with an Accuracy of 0.872 and the highest F1 Score of 0.622. Random Forest had a marginally higher Precision (0.772 vs 0.767) but a slightly lower Recall (0.500 vs 0.523) compared to XGBoost.

Best-Performing Model

Based on the F1 Score, which is a crucial metric for imbalanced datasets as it balances Precision and Recall, the **XGBoost Classifier** is identified as the best-performing model. It achieved an F1 Score of 0.622, indicating a strong balance between correctly identifying high-risk individuals (Recall) and minimizing false alarms (Precision).

Insights from ROC Curves and Confusion Matrices

- **ROC Curves**: The ROC curves for Random Forest and XGBoost Classifier demonstrate superior performance, with AUC scores significantly higher than the other models (likely above 0.85-0.90 based on the F1 scores). A high AUC indicates that these models are better at distinguishing between the two classes (high-risk vs. not high-risk).
  - For XGBoost Classifier, its ROC curve would be closer to the top-left corner, signifying a higher true positive rate for a given false positive rate, implying robust discrimination capability.
- **Confusion Matrices**: Analyzing the Confusion Matrix for the best-performing XGBoost Classifier reveals its strengths:
  - It correctly identified a substantial number of **True Negatives** (individuals correctly classified as not high-risk), contributing to its high Accuracy.
  - It also effectively identified a good proportion of **True Positives** (individuals correctly classified as high-risk), reflected in its Recall. The balance between True Positives and False Positives (indicated by its Precision) is better than simpler models, suggesting it minimizes misclassifying non-high-risk individuals as high-risk.

Feature Importance

Based on the SHAP summary plot for the XGBoost model and the feature importance plots for both XGBoost and Random Forest classifiers, we can identify several critical features influencing the prediction of 'high_risk_flag' (dopamine overload risk).

SHAP Values for XGBoost Model:

The SHAP summary plot provides insights into how each feature impacts the model's output for individual predictions. The features are ordered by importance, and the colors represent the feature value (red for high, blue for low). High SHAP values indicate a stronger impact on pushing the prediction towards a higher risk (class 1).

- **`num__digital_dependence_score`**: This feature consistently appears as one of the most important. High values (red dots to the right) tend to increase the SHAP value, pushing the model to predict a higher `high_risk_flag`. This aligns with intuition: individuals with higher digital dependence are more likely to experience dopamine overload.
- **`num__mental_health_load`**: Also highly influential. Higher mental health load (sum of anxiety, depression, stress) strongly correlates with increased risk, as shown by higher SHAP values for higher feature values.
- **`num__device_hours_per_day`** and **`num__notifications_per_day`**: These features directly relate to digital usage. Higher values for these features generally lead to increased SHAP values, contributing to a higher predicted risk. This suggests that extensive device usage and constant notifications are significant risk factors.
- **`num__stress_level`**, **`num__anxiety_score`**, **`num__depression_score`**: As components of `mental_health_load`, these individual scores also show strong influence, with higher scores pushing towards higher risk.
- **`num__sleep_quality`**: Lower sleep quality appears to be associated with higher risk. The SHAP plot often shows blue dots (lower quality) on the positive side of the SHAP value axis.

Comparison of Feature Importance (XGBoost vs. Random Forest):

Looking at the bar plots for feature importance:

- **Common Top Features**: Both models highlight `digital_dependence_score`, `mental_health_load`, `stress_level`, `device_hours_per_day`, `notifications_per_day`, `anxiety_score`, and `depression_score` as consistently important. This strong agreement across different tree-based models reinforces the significance of these factors.
- **Slight Differences**: While the top features are largely similar, their exact ranking and relative importance can vary slightly between XGBoost and Random Forest. For instance, XGBoost might give slightly more weight to `digital_dependence_score` or `mental_health_load`, while Random Forest might emphasize other related usage metrics. This indicates that both models capture the underlying relationships well, but with nuanced perspectives.

Real-World Implications:

The identified key features offer crucial insights into the mechanisms behind digital dependence and dopamine overload:

- **`digital_dependence_score`**: This feature's prominence suggests that a direct measure of digital dependence is the strongest indicator of risk. High scores here mean individuals are heavily reliant on digital devices, which directly correlates with the `high_risk_flag` for dopamine overload.

- **`mental_health_load` (and its components like `stress_level`, `anxiety_score`, `depression_score`)**: The strong influence of these mental health indicators implies a bidirectional relationship. Individuals with pre-existing or heightened mental health challenges might be more susceptible to unhealthy digital behaviors, or excessive digital engagement could exacerbate these conditions, leading to dopamine overload. This highlights the importance of mental well-being in mitigating digital risks.
- **`device_hours_per_day` and `notifications_per_day`**: These features represent direct measures of digital consumption and engagement. High values signify prolonged exposure and constant digital stimulation, which are primary drivers of dopamine release and, consequently, overload. This points to the need for managing screen time and reducing notification frequency.
- **`phone_unlocks`**: Similar to device hours and notifications, frequent phone unlocks indicate compulsive checking behavior, a hallmark of digital dependence.
- **`sleep_quality`**: Poor sleep quality is a known consequence of excessive screen time, especially before bed. Its importance here suggests that disrupted sleep patterns are not just symptoms but also contributors to the overall risk of digital dependence and its associated mental health issues.

In essence, the models confirm that extensive and habitual digital engagement, coupled with underlying mental health vulnerabilities, are strong predictors of dopamine overload risk. These findings can inform intervention strategies focusing on digital detox, mental health support, and promoting healthier digital habits.

Regression Model

The model achieved the following metrics:

- Mean Squared Error (MSE): 5.40
- R-squared (R2) Score: 0.97

True vs. Predicted Digital Dependence Score Plot

The "True vs. Predicted Digital Dependence Score" plot visually compares the actual `digital_dependence_score` values against the model's predictions. The scatter plot shows a strong linear relationship between the true and predicted values, with points closely clustered around the "Perfect Prediction Line" (red dashed line). An R-squared value of 0.97 signifies that approximately 97% of the variance in the `digital_dependence_score` can be explained by the model, which is an excellent fit. The tight clustering of points around the perfect prediction line indicates that the model is generally making accurate predictions across the range of scores, with minimal bias.

Residual Plot for Linear Regression (Ridge)

The "Residual Plot for Linear Regression (Ridge)" displays the residuals (the difference between true and predicted values) against the predicted values. A well-fitting linear regression model should ideally show a random scatter of residuals around the zero line, with no discernible patterns. In this plot, the residuals are mostly scattered around the horizontal zero line. This random distribution suggests that the model's assumptions (e.g., linearity, homoscedasticity) are largely met and that there are no obvious systematic errors in the predictions that would indicate a poor fit or missing variables. The absence of patterns (like a cone shape or a curve) indicates that the model performs consistently across the range of predicted values, further supporting its robustness.

Conclusion

This project set out to develop a machine learning–based approach for predicting dopamine overload risk and digital dependence using digital lifestyle and mental health indicators. By implementing a complete end-to-end pipeline, from data exploration and preprocessing to feature engineering and model evaluation, the project demonstrated that meaningful and actionable patterns can be learned from such data.

The outcomes of the project emphasize that digital dependence and associated risks are influenced by a combination of behavioral intensity, mental health burden, and lifestyle factors, rather than by device usage alone. This integrated perspective strengthens the practical relevance of the project and highlights the value of combining psychological and behavioral features in predictive systems related to digital well-being.

While the project achieved strong predictive performance, it also revealed several opportunities for improvement. Future enhancements could include further model optimization, exploration of more advanced learning techniques, and the use of larger and more diverse datasets, particularly longitudinal or real-time data, to better capture behavioral dynamics over time. Additionally, deeper interpretability and causal analysis would improve trust and usability, especially in applications aimed at early intervention or personalized feedback.

In summary, this project demonstrates the potential of machine learning as a practical tool for assessing digital dependence and dopamine overload risk, while also providing a clear foundation for future extensions toward more robust, interpretable, and real-world–deployable systems.

References:

1. Digital Lifestyle Benchmark dataset
2. Dopamine_Overload_Prediction.ipynb