College of Computer and Information Sciences
Information Technology Department

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

The Logbook

Primary Dataset

Group 1

Supervised by:

Dr.Khulood Alyahya

2024-2025

Table of Contents

1. Primary Overview 3

2. Data Collection and Preparation 3

3. Model Building 5

4. HMI Calculation 7

5. Model Evaluation 9

6. Model Performance Results 10

7. Performance Visualization 13

8. Conclusion of Model Evaluation 15

9. Challenges and Solutions 15

10. References 17

1. Primary Overview

The goal of this project is to develop a predictive model to estimate the Mental Health Index (MHI) based on multiple factors associated with social media usage and its effects on mental health. The dataset includes demographic information and survey responses that capture behaviors related to self-esteem, anxiety, insomnia, FOMO (Fear of Missing Out), and social media habits. The challenge was to build a reliable model that accurately predicts the MHI score using these features.

Data Source:
The dataset used in Primary sourse is collected from a survey on social media usage and its psychological effects. It consists of 500+ rows and includes columns such as age, gender, social media usage patterns, and responses to mental health-related questions like self-esteem, anxiety, insomnia, and FOMO.

2. Data Collection and Preparation

The dataset contains both **categorical** (e.g., gender, employment status) and **numerical** features (e.g., age, hours spent on social media). The target variable, **Mental Health Index (MHI)**, is a continuous variable derived from responses to multiple questions about mental health indicators such as **self-esteem**, **social anxiety**, **FOMO**, etc. These components are weighted to form the MHI score.

- **Columns Removed:**
    - The columns **"Do you use social media applications?"** and **"What social media platforms do you use?"**were dropped as they were irrelevant to the model's purpose or duplicated the information found in other columns. This helps reduce complexity and ensures we only work with relevant features.

- **Feature Encoding:**
    - **Label Encoding**: Applied to the **"Gender"** column to convert the categorical labels ("Male" and "Female") into numeric labels (0 and 1). This was necessary because most machine learning models cannot handle categorical data directly.

    - **Why Label Encoding?**: Label Encoding was used for binary variables (e.g., gender), where the ordering of categories is not important, and we only need to represent them numerically.

  - **One-Hot Encoding**: Applied to multi-category features, including **'the age:'**, **'Area:'**, **'Current educational level:'**, **'marital status:'**, **'Employment status:'**, and **'What app do you use the most?'**. These features have more than two categories and need to be transformed into separate binary columns (one for each category).

    - **Why One-Hot Encoding?**: One-Hot Encoding was chosen because these features are nominal (no inherent order), and One-Hot Encoding creates separate columns for each category, thus avoiding assumptions about the ordinal nature of the categories.

- **Data Transformation:**

  - After encoding the categorical columns, the **original columns** were removed, and the new encoded columns were concatenated into the DataFrame. This ensures that the dataset is now in a format suitable for machine learning models.

- **Normalization:**

Z-Score Normalization was applied to specific survey questions related to self-esteem, anxiety, FOMO, and other social media-related behaviors to standardize the features for regression modeling. The normalization was done using StandardScaler from scikit-learn, ensuring that the selected columns have a mean of 0 and a standard deviation of 1.

This transformation is crucial for regression models because it:

- **Prevents features with larger scales from dominating the model.**
- **Ensures that all features contribute equally to the model's performance.**

The columns that were normalized include:

- **Self-esteem-related behaviors (e.g., "Does the number of likes or comments you get on your posts affect you?")**
- **Social anxiety-related behaviors (e.g., "Do you feel anxious or stressed after reading negative comments on your posts?")**

- **FOMO-related behaviors (e.g., "Are you worried about missing out on important information or events when you're not using social media?")**
- **General social media usage patterns (e.g., "Do you use social media right before going to sleep?")**

By normalizing these features, we ensure that they are all on a comparable scale, improving the performance of regression models like Linear Regression, Random Forest, and Gradient Boosting.

3. Model Building

Model Selection:
After analyzing the dataset and understanding the relationships between social media usage and mental health indicators, we chose a variety of models to predict the **Mental Health Index (MHI)**. Each model was selected for its ability to handle different types of data and relationships.

1. Linear Regression (Baseline Model)

Linear Regression was chosen as the baseline due to its simplicity and interpretability. It assumes a linear relationship between features (e.g., age, hours spent on social media) and the target variable (MHI). It fits a straight line to the data by optimizing coefficients to minimize the residual sum of squares. While effective for simple relationships, it may not capture complex, non-linear patterns but provides a useful starting point for comparison with more advanced models.

2. Random Forest Regressor

Random Forest is an ensemble model that combines multiple decision trees, each trained on a random data subset. The final prediction is the average of all trees. It is particularly good at capturing complex, non-linear relationships and identifying influential features like social media usage and FOMO. This model is effective for handling large datasets with numerous features and reduces overfitting.

3. Gradient Boosting Regressor

Gradient Boosting builds decision trees sequentially, with each tree correcting errors made by the previous one. This iterative approach allows it to capture subtle patterns and complex relationships. It outperformed other models in this project, achieving the highest **R^2** score, and is well-suited for datasets with intricate feature relationships like those between social media habits and mental health outcomes.

4. Support Vector Regressor (SVR)

SVR uses a kernel trick to map data into higher-dimensional spaces, capturing non-linear relationships that linear models can't represent. It defines a margin of tolerance and fits a regression line within this margin, making it powerful for handling complex patterns in social media usage and mental health scores without assuming a specific data form.

5. K-Nearest Neighbors (KNN)

KNN is a simple model that predicts the target variable based on the values of its k nearest neighbors in the feature space, using distance metrics like Euclidean distance. It is effective for detecting patterns in groups with similar behaviors, but can be computationally expensive as the dataset grows and its performance may degrade with sparse or irrelevant data.

Why These Models Were Chosen: The combination of **Linear Regression**, **Random Forest**, **Gradient Boosting**, **SVR**, and **KNN** allows us to explore a variety of approaches for predicting **MHI**. The baseline model of **Linear Regression** provides a simple and interpretable result, while the more complex models like **Random Forest** and **Gradient Boosting** allow us to capture non-linear relationships and interactions between features. **SVR** and **KNN** are included for their ability to model subtle and complex patterns without making assumptions about the form of the data.

The choice of models reflects the goal of identifying the most appropriate one for this specific dataset, with the **Gradient Boosting Regressor** emerging as the most accurate in terms of prediction performance.

Train-Test Split:

To ensure the model generalizes well to new, unseen data and to prevent overfitting, the dataset was divided into **training**and **testing** sets. The training set is used to train the model, while the test set serves as a holdout dataset to evaluate the model's performance.

- **Independent Variables (Features)**: These include demographic information (age, gender) and behavior-related responses (social media usage, FOMO, etc.), encoded and normalized as discussed.
- **Dependent Variable (Target)**: The **Mental Health Index (MHI)**, which is the target that models aim to predict.

The **train-test split** was set at 70% for training data and 30% for testing data. This separation of data helps evaluate the model's performance on unseen data, mimicking real-world scenarios where models need to generalize beyond training examples.

4. HMI Calculation

4.1 Mental Health Index (MHI) Weights Distribution with Brain Functions and Cognitive Processes:

The weights for the components of the **Mental Health Index (MHI)** were assigned based on input from a **psychologist**who holds a **master's degree in psychology** and works as a **consultant at a hospital**. This expert helped ensure that the weights accurately represent the psychological and cognitive importance of each factor, as informed by psychological literature and clinical experience. The MHI was computed by taking the **weighted average** of the components, based on the responses in the survey.

The following factors were weighted according to their impact on mental health:

- **Self-Esteem:** 35%
- **Social Anxiety:** 25%
- **Insomnia:** 20%
- **Fear of Missing Out (FOMO):** 15%
- **Shorter Attention Span:** 5%

Each factor’s score was calculated by averaging the responses to the relevant survey questions. The final MHI score was computed as a weighted sum of these individual scores.

4.2 Explanation of Distribution:

- **Self-Esteem** is the most important factor (35%) because it impacts emotional regulation and overall mental well-being.
- **Social Anxiety** (25%) affects emotional responses, and **Insomnia** (20%) impacts cognitive function and emotional regulation.
- **FOMO** (15%) contributes to stress, while **Attention Span** (5%) has an indirect effect on mental health.

4.3 Calculation Process

For each factor (like **Self-Esteem**, **Social Anxiety**, etc.), the **average score** of the relevant questions is computed first. Then, the weighted average of these components is taken to compute the final **MHI**.

Step-by-step MHI Calculation:

1. **Grouping Questions**: The questions related to each factor (e.g., **Self-Esteem**, **Social Anxiety**) are grouped together in a dictionary (columns_mapping).
   - For example, columns_mapping["Self-Esteem"] contains the questions related to **Self-Esteem**.
2. **Calculating Average for Each Factor**:
   - For each factor (e.g., **Self-Esteem**), the code computes the mean of the selected questions in the corresponding group.
   - data["Self-Esteem"] = data[columns_mapping["Self-Esteem"]].mean(axis=1) calculates the mean score for each row (respondent) across the **Self-Esteem** related questions.
3. **Weighted MHI Calculation**: Once all the factors are averaged, the **MHI** is calculated by applying the weights from the weights dictionary to each factor's average score.

- This line multiplies each factor's average by its respective weight and adds the results to get the final MHI score.
- The `weights` dictionary defines how much each component contributes to the final score.

The final MHI score is a weighted sum of all these components.

5. Model Evaluation

5.1 Training the Models

After preprocessing the data, the models were trained using the training dataset. The performance of each model was then evaluated on the test dataset to ensure they generalize well to unseen data.

As we mentioned the following models were evaluated:

1. Linear Regression (Baseline Model)
2. Random Forest Regressor
3. Gradient Boosting Regressor
4. Support Vector Regressor (SVR)
5. K-Nearest Neighbors (KNN)

Training Process: For each model, the process involved:

- Splitting the data into training and test sets (70% training, 30% testing).
- Fitting the model to the training data using the .fit() method.
- Making predictions on the test data using the .predict() method.

5.2 Performance Evaluation Metrics

For evaluating model performance, the following metrics were used:

1. **Mean Squared Error (MSE)**: Measures the average squared difference between the predicted and actual values. A lower MSE indicates better performance.

2. **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual values. Like MSE, lower values are better.

3. **Root Mean Squared Error (RMSE)**: The square root of MSE, providing an interpretation of prediction error. It gives more weight to larger errors.

4. **R² Score**: Measures how well the model explains the variance in the target variable. An **R²** score closer to 1 indicates a better fit.

The models were evaluated on these metrics to determine how well they performed on predicting the **Mental Health Index (MHI)**.

6. Model Performance Results

6.1. Null Model

The **Null Model** serves as a baseline, predicting the mean value of the target variable (MHI) for all test data. This helps us understand how well our models perform compared to a simple model that doesn't use any features.

The **Null Model's Performance** was as follows:

- **MSE**: 0.3554
- **MAE**: 0.4900
- **RMSE**: 0.5961
- **R² Score**: -0.0094 (indicating poor performance)

A **negative R² score** indicates that the Null Model performed worse than just predicting the mean value of the MHI, which is expected since it's not learning anything from the data.

6.2. Baseline Model - Linear Regression

Linear Regression is considered the baseline model. This model assumes a **linear relationship** between the features and the target variable. It is **simple** but interpretable and provides a good starting point for comparison.

The **Linear Regression Model's Performance**:

- **MSE**: 0.3626

- **MAE**: 0.4900
- **RMSE**: 0.6022
- **R² Score**: -0.0300 (similar to the Null Model, suggesting it couldn't capture complex patterns)

Interestingly, the **Linear Regression Model** performed similarly to the **Null Model**. Despite the simplicity of Linear Regression, it had an **R² score** close to 0, suggesting it couldn't capture complex patterns in the data.

6.3. Random Forest Regressor

Random Forest Regressor is an **ensemble learning model** that uses multiple decision trees to predict outcomes. By aggregating the results of many trees, it reduces the risk of overfitting and can handle non-linear relationships.

The **Random Forest Regressor's Performance**:

- **MSE**: 0.3946
- **MAE**: 0.5081
- **RMSE**: 0.6282
- **R² Score**: -0.1208 (indicating poor performance)

Random Forest performed poorly with a **negative R² score** and relatively higher **MSE** and **RMSE**, suggesting it failed to capture useful patterns in the data.

6.4. Gradient Boosting Regressor

Gradient Boosting builds decision trees sequentially, each tree correcting errors made by the previous one. This iterative process allows **Gradient Boosting** to gradually improve its predictions and handle more complex relationships in the data.

The **Gradient Boosting Regressor's Performance**:

- **MSE**: 0.3601
- **MAE**: 0.4860
- **RMSE**: 0.6001

- **R² Score**: -0.0229 (outperformed other models but still underperformed)

Despite having a **negative R² score**, **Gradient Boosting** outperformed **Random Forest** and **Linear Regression** in terms of **MSE**, **MAE**, and **RMSE**, though it still did not explain the majority of variance in the data.

6.5. Support Vector Regressor (SVR)

The **Support Vector Regressor (SVR)** uses the **kernel trick** to map data to a higher-dimensional space, enabling it to model **non-linear relationships** between features and the target variable.

The **SVR's Performance**:

- **MSE**: 0.3692
- **MAE**: 0.5019
- **RMSE**: 0.6076
- **R² Score**: -0.0487 (moderate performance but still negative)

SVR showed **moderate performance** but still had a **negative R² score**, suggesting that it did not capture meaningful patterns in the data.

6.6. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and intuitive model that classifies the target variable based on the **k nearest neighbors** in the feature space. This model was included to capture any underlying groupings in the data.

The **KNN's Performance**:

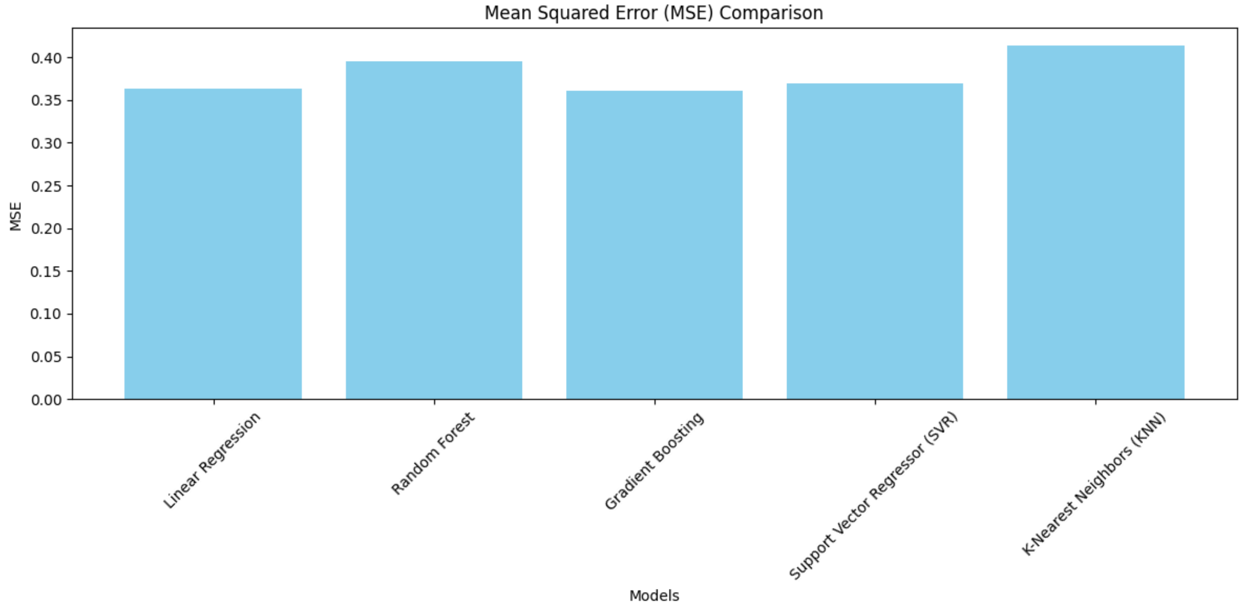
- **MSE**: 0.4134
- **MAE**: 0.5332
- **RMSE**: 0.6430
- **R² Score**: -0.1743 (poorest performance of all models)

KNN performed the **worst** out of all models, with the **highest MSE** and **RMSE**, and a **very negative R² score**, indicating it struggled to model the relationships between features and the MHI.

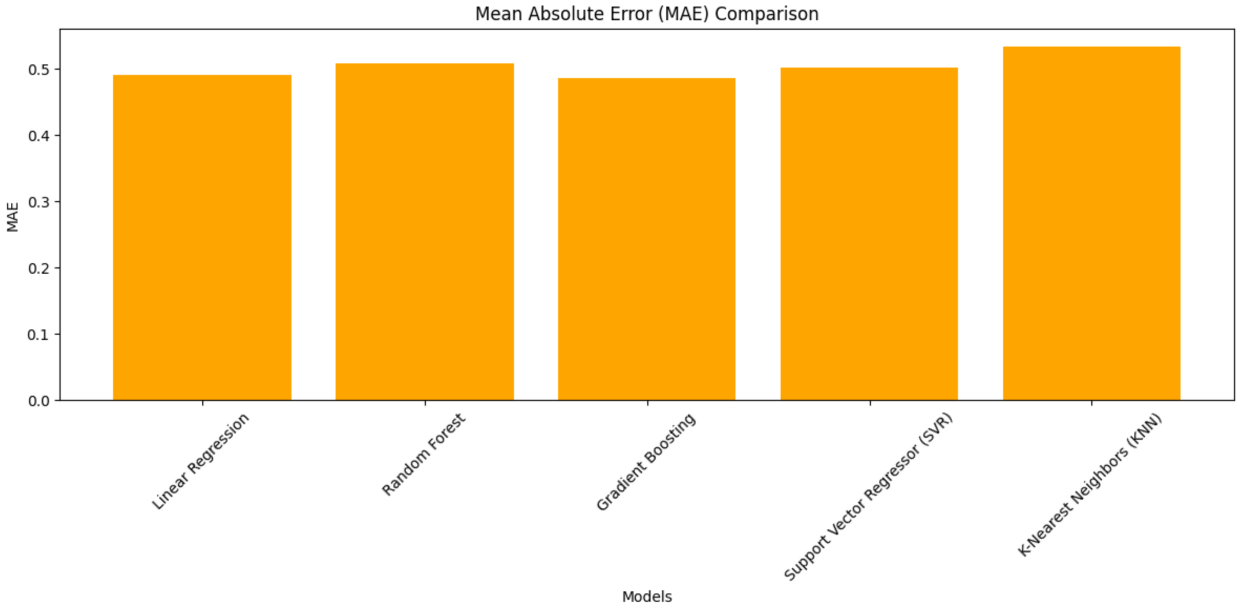
7. Performance Visualization

To better understand the performance of each model, we visualized the comparison of key evaluation metrics:

1. **MSE** Comparison:
   - **Linear Regression** performed similarly to the **Null Model**, with **higher MSE** than other models.
   - **Gradient Boosting** had a slightly lower **MSE**, but none of the models performed exceptionally well.

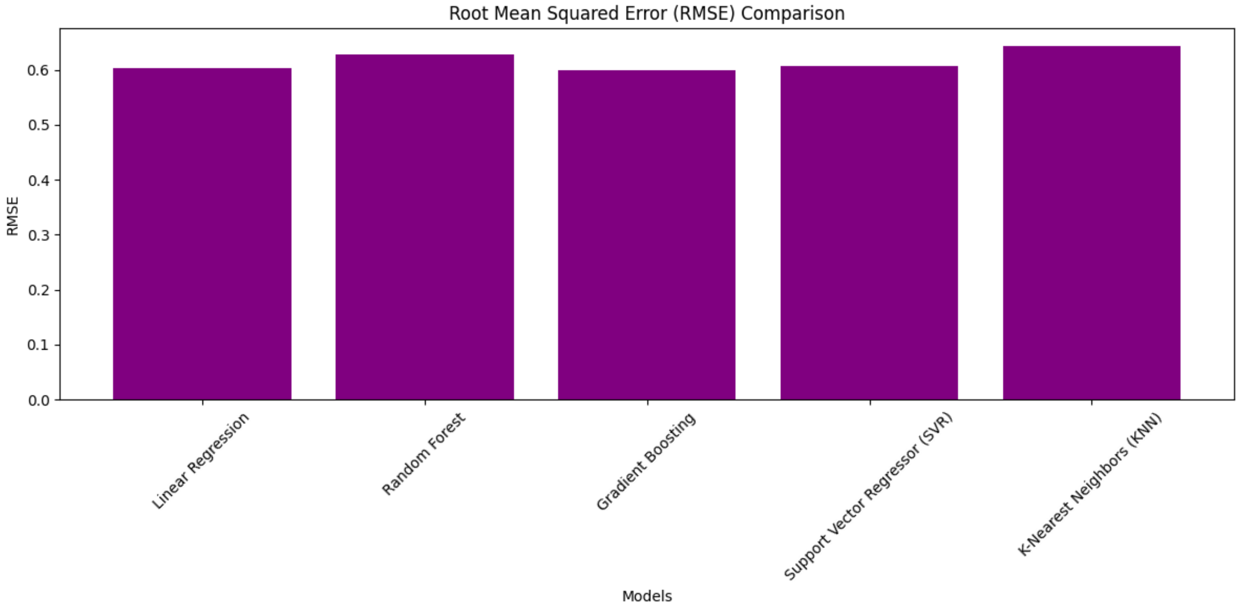


2. **MAE** Comparison:
   - **Gradient Boosting** had the lowest **MAE**, but all models still had significant errors, indicating that predictions were off by a considerable amount.



3. **RMSE** Comparison:

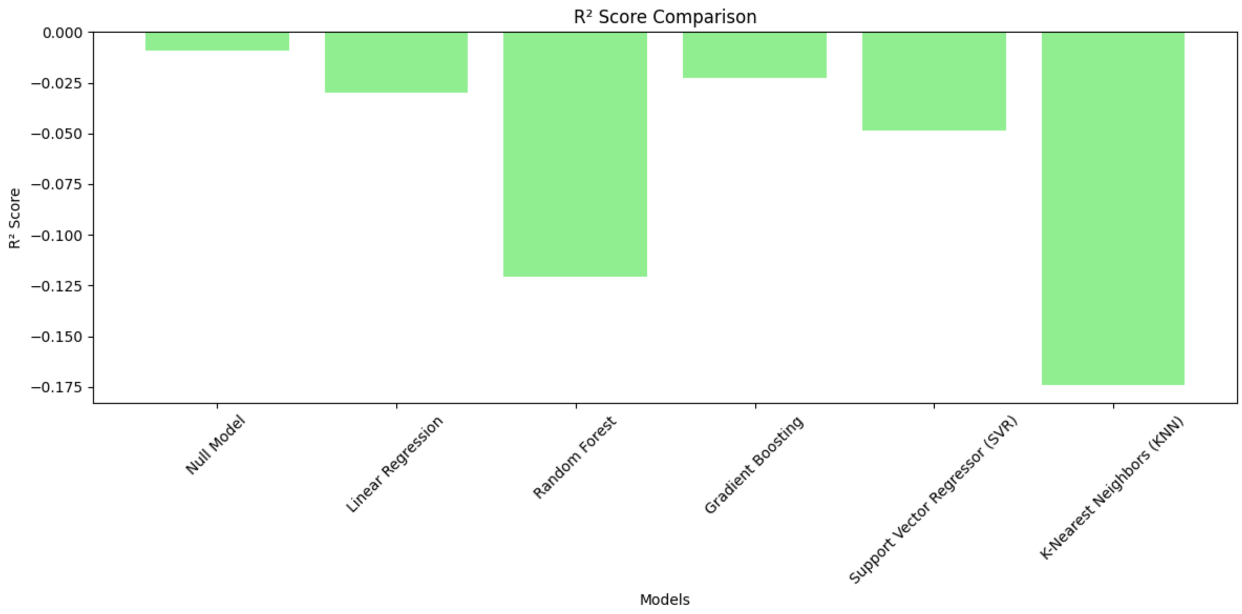
   ○ Similar to **MSE**, **Gradient Boosting** had the lowest **RMSE**, followed by **Random Forest** and **Linear Regression**.



4. **R² Score** Comparison:

   ○ **Linear Regression** had the **highest R² score**, although it was still negative, indicating poor performance.

- **KNN** had the **lowest R² score**, suggesting it performed the worst.



8. Conclusion of Model Evaluation

- **Linear Regression** proved to be the best performing model, even though it couldn't fully capture complex patterns in the data. It has the highest R² score and lowest error metrics compared to other models.
- **Gradient Boosting** and **SVR** showed promise with relatively higher performance but still fell short due to negative R² scores.
- **Random Forest** and **KNN** underperformed, particularly with negative R² scores, indicating that they did not effectively capture meaningful patterns in the dataset.

9. Challenges and Solutions

Data Issues:

- **Handling Irrelevant Columns:**
  In the code, two columns, **"Do you use social media applications?"** and **"What social media platforms do you use?"**, were dropped from the dataset. These columns were likely deemed irrelevant, but it was a challenge to ensure that only the most relevant data remained for feature selection.

- **Solution:** The irrelevant columns were removed using the drop() function to focus on features that contribute directly to the prediction of the target variable.

Modeling Challenges:

- **Feature Encoding:**
  Some categorical features (e.g., **Gender**, **Age**, **Employment status**) needed to be encoded for the models to process them. Converting categorical data into a numerical format using **Label Encoding** and **One-Hot Encoding**, especially with handling many unique categories or ensuring that no information is lost during transformation.

  - **Solution: Label Encoding** was used for binary categorical data (e.g., **Gender**), and **One-Hot Encoding** was applied to multi-class categorical data (e.g., **Employment status**, **Area**, etc.). This ensured that categorical data could be used in model training while preserving all relevant information.

10. References

Self-Esteem

1. **Eisenberger, N. I., & Cole, S. W.** (2012). "Social neuroscience and health: Neurophysiological mechanisms linking social ties with physical health." *Nature Neuroscience*.
   Link
2. **Orth, U., et al.** (2012). "Self-esteem development from young adulthood to old age: A cohort-sequential longitudinal study." *Journal of Personality and Social Psychology*.
   Link

Social Anxiety

1. **Etkin, A., & Wager, T. D.** (2007). "Functional neuroimaging of anxiety: A meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia." *American Journal of Psychiatry*.
   Link
2. **Stein, M. B., et al.** (2001). "Neurobiology of generalized anxiety disorder: A review." *CNS Spectrum*.
   Link

Insomnia

1. **Saper, C. B., et al.** (2005). "The sleep switch: Hypothalamic control of sleep and wakefulness." *Trends in Neurosciences*.
   Link
2. **Walker, M. P.** (2008). "Cognitive consequences of sleep deprivation." *Sleep Medicine Clinics*.
   Link

Fear of Missing Out (FOMO)

1. **Meshi, D., et al.** (2013). "Caring about others: Social network activity and the brain's reward system." *Journal of Cognitive Neuroscience*.
   Link

2. **Przybylski, A. K., et al.** (2013). "Motivational, emotional, and behavioral correlates of fear of missing out." *Computers in Human Behavior*.
   Link

Shorter Attention Span

1. **Raichle, M. E.** (2015). "The brain's default mode network." *Annual Review of Neuroscience*.
   Link

2. **Ophir, E., et al.** (2009). "Cognitive control in media multitaskers." *Proceedings of the National Academy of Sciences (PNAS)*.
   Link