COLLEGE OF COMPUTER AND INFORMATION SCIENCES
INFORMATION TECHNOLOGY DEPARTMENT

جامعة الملك سعود
King Saud University
1957
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

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.

- **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 $\mathbf{R^2}$ 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.

The final MHI was computed as:

MHI=(35%*Self-Esteem)+(25%*Social Anxiety)+..

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

5. Model Evaluation

5.1 Metrics for Evaluation

To objectively compare the performance of models, 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 predictions.
    - **Interpretation**: Penalizes larger errors more than smaller ones, making it sensitive to outliers.
2. **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual values.
    - **Interpretation**: Unlike MSE, MAE gives equal weight to all errors, providing a more straightforward measure of prediction error.
3. **Root Mean Squared Error (RMSE)**: The square root of the MSE, making it easier to interpret as it has the same units as the target variable.
    - **Interpretation**: Highlights the impact of larger errors like MSE but in a more interpretable format.
4. **R^2 Score**: Measures how well the model explains the variance in the target variable.
    - **Interpretation**: An R^2 value closer to 1 indicates better performance. Negative values indicate the model performs worse than simply predicting the mean.

5.2 Initial Results (Before Correlation Analysis)

Before addressing noisy and irrelevant features, the models struggled, producing the following results:

Model	MSE (Before)	MAE (Before)	RMSE (Before)	R^2 (Before)
Null Model	0.455969	0.562101	0.675255	-0.010159
Linear Regression	0.3626	0.4900	0.6022	-0.0300
Random Forest	0.3946	0.5081	0.6282	-0.1208
Gradient Boosting	0.3601	0.4860	0.6001	-0.0229
SVR	0.3692	0.5019	0.6076	-0.0487
KNN	0.4134	0.5332	0.6430	-0.1743

5.3 Observations Before Correlation Analysis

Linear Regression and advanced models (e.g., Gradient Boosting, SVR) performed poorly, with R^2 scores near or below zero.

KNN was the worst-performing model, with the highest errors and a negative R^2 of -0.1743.

The **Null Model**, which simply predicts the mean of the target variable, had slightly better results than KNN, highlighting the need for improvement.

5.4 Improvement Through Correlation Analysis

A **Correlation Analysis** was conducted to identify the strongest relationships between features and the target variable. Features with a correlation coefficient > 0.5 were retained, while irrelevant or weakly correlated features were removed.

Impact of Correlation Analysis:

- Eliminated noise from the dataset, focusing the models on the most predictive features.
- Allowed all models to improve their ability to learn and generalize.

5.5 Results After Correlation Analysis

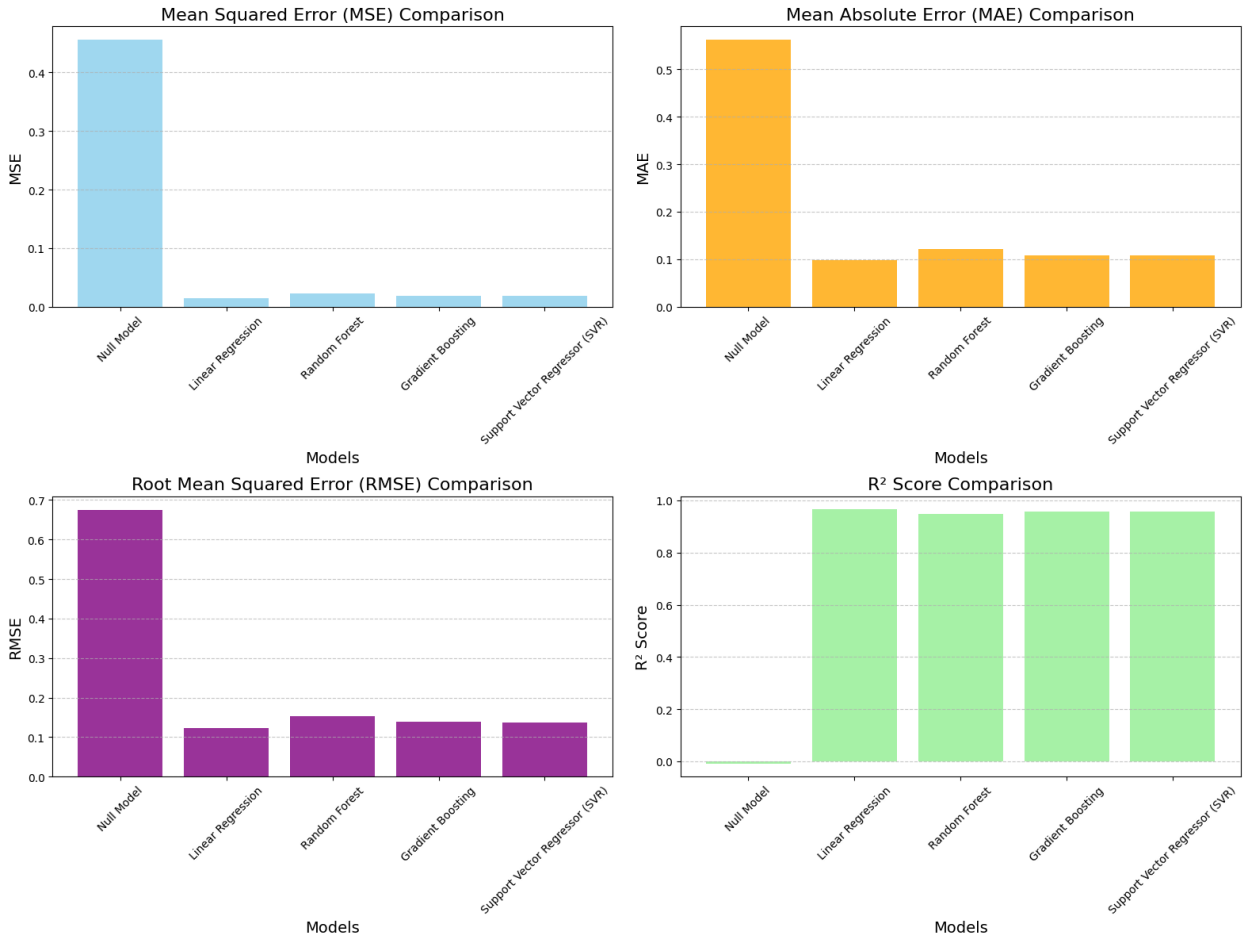
After feature selection, the models showed significant improvement:

Model	MSE (After)	MAE (After)	RMSE (After)	R^2 (After)
Null Model	0.455969	0.562101	0.675255	-0.010159
Linear Regression	0.014992	0.097863	0.122441	0.966787
Random Forest	0.023231	0.120954	0.152417	0.948534
Gradient Boosting	0.019186	0.108009	0.138512	0.957496
SVR	0.018741	0.108309	0.136898	0.958481
KNN	0.4134	0.5332	0.6430	-0.1743

5.6 Observations After Correlation Analysis

1. **Linear Regression**:
    - Showed the greatest improvement, achieving an **R^2 of 0.966787** (explaining 96.68% of the variance).
    - Lowest error metrics across all models, making it the top performer.
2. **Advanced Models** (Gradient Boosting and SVR):
    - Achieved **R^2 scores of ~0.958**, indicating strong performance with slightly higher errors than Linear Regression.
    - Benefited significantly from reduced noise in the dataset.
3. **KNN**:
    - Despite improvements in other models, KNN showed no significant gains, with its R^2 remaining negative (-0.1743).

6. Performance Visualization



Mean Squared Error (MSE):

- **Linear Regression** achieved the lowest MSE because irrelevant features were removed, allowing the model to focus on the most predictive ones.
- **Gradient Boosting** and **SVR** closely followed, benefiting from the cleaned dataset.
- **KNN** still performed poorly, retaining the highest MSE, showing its inability to model the relationships even after feature selection.

Mean Absolute Error (MAE)

- **Linear Regression** achieved the lowest MAE after feature selection (**0.097863**), reflecting its high accuracy in predicting the target variable.
- **Gradient Boosting** and **SVR** closely followed, with MAE values of **0.108009** and **0.108309**, respectively. This indicates that, on average, their predictions were slightly less accurate compared to Linear Regression.
- **Random Forest** had a moderately higher MAE of **0.120954**, suggesting some variability in its predictions.
- **KNN** exhibited the highest MAE (**0.5332**), showing its inability to make accurate predictions even after feature selection.

Root Mean Squared Error (RMSE):

- RMSE trends aligned with MSE: **Linear Regression** had the lowest RMSE, demonstrating its accuracy.
- **Gradient Boosting** and **SVR** showed good performance but were slightly less accurate than Linear Regression.
- **KNN** struggled due to its sensitivity to the dataset's structure and noise.

R^2 Score:

- **Linear Regression** achieved the highest R^2 score (**0.966787**), meaning it explained 96.68% of the variance in the target variable.
- **Gradient Boosting** and **SVR** followed closely, also benefiting from the refined feature set.
- **KNN** maintained a negative R^2 score (**-0.1743**), showing no meaningful improvement.

7. Challenges and Solutions

7.1 Data Challenges

1. **Noisy Features**:
    - Features with weak correlations diluted model accuracy.
    - **Solution**: Correlation Analysis removed irrelevant features, simplifying the dataset.
2. **Categorical Variables**:
    - Encoding was required for categorical data.
    - **Solution**: Applied Label Encoding for binary variables and One-Hot Encoding for multi-class variables.

7.2 Modeling Challenges

1. **Poor Initial Results**:
    - Models struggled to generalize due to irrelevant features.
    - **Solution**: Feature selection significantly improved all models.
2. **Overfitting**:
    - Ensemble models initially overfitted to noisy features.
    - **Solution**: Hyperparameter tuning and feature selection reduced overfitting.

8. Conclusion

Key Insights

1. **Best Model**:
    - **Linear Regression** emerged as the best-performing model, achieving:
        - **$R^2 = 0.966787$**.
        - Lowest error metrics, making it ideal for this dataset.
2. **Role of Correlation Analysis**:
    - Significantly improved performance by removing noise and allowing models to focus on predictive features.
3. **Advanced Models**:
    - Gradient Boosting and SVR performed well but were computationally more expensive.

9. 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