Sri Lanka Institute of Information

Technology



Final Report

|  |  |
| --- | --- |
| **2025-Y2-S1-MLB-B9G2-03** | |
| **Web-based Hotel Reservation System for Tourists** | |
| IT24102544 | Adhikari A.Y.S |
| IT24102483 | Premathilaka H.P.S.M |
| IT24102482 | Jayathilaka D.L.T.S |
| IT24102552 | Hannadige L.H.D |
| IT24102476 | Mohomed N.M.N |
| IT24102466 | Hariluxshun.R |

**AIML | IT2011**

B.Sc. (Hons) in Information Technology

**1. Introduction and Problem Statement**

Mental health disorders, particularly depression, are prevalent worldwide, affecting individuals across various demographics. Early detection and intervention are crucial for effective treatment. Traditional diagnostic methods often rely on clinical assessments, which can be subjective and time-consuming. With the advent of machine learning, there is an opportunity to develop predictive models that can assist in identifying individuals at risk of depression based on their responses to lifestyle and socio-economic factors.

This project aims to build and evaluate machine learning models to predict depression indicators, utilizing the dataset provided by Anthony Therrien. The objectives are to:

* Develop predictive models using various machine learning algorithms.
* Evaluate model performance using appropriate metrics.
* Analyze the ethical implications and potential biases in the dataset and models.

**2. Dataset Description**

The dataset, sourced from [Kaggle](https://www.kaggle.com/datasets/anthonytherrien/depression-dataset/data), contains information on individuals' lifestyle and socio-economic factors. It includes features such as:

* Age, gender, marital status
* Employment status, education level
* Physical activity, sleep patterns
* Social media usage, internet addiction
* Depression indicator (target variable)

The dataset comprises more than 400000 records with 16 features. While the dataset provides valuable insights, it may contain biases due to overrepresentation of certain demographics, such as young adults and individuals from specific geographic regions. These biases can impact the generalizability of the models.

**3. Preprocessing & EDA**

3.1 Data Cleaning

* Loaded the dataset (depression\_data.csv) and confirmed no missing values.
* Removed duplicate or redundant columns.
* Checked for outliers in numeric features (Age, Number of Children, Income) using the IQR method and removed extreme values.

3.2 Feature Encoding

* Ordinal features (e.g., Education Level, Sleep Patterns, Physical Activity Level) were encoded with meaningful orderings.
* Binary/categorical features (e.g., History of Mental Illness, Family History of Depression) were one-hot encoded.
* All preprocessing was applied using ColumnTransformer for consistency.

3.3 Feature Scaling

* Numeric features were standardized using StandardScaler to ensure comparable ranges across models.
* Scaling was mainly applied for models sensitive to feature magnitude (e.g., MLP, Logistic Regression).

3.4 Class Balancing

* Target variable was slightly imbalanced.
* Balanced dataset using strategic sampling: selected the most informative majority-class samples using Random Forest probabilities to match minority class size.
* Resulting dataset had equal representation of depression and non-depression cases.

3.5 Life Factor Index (LFI)

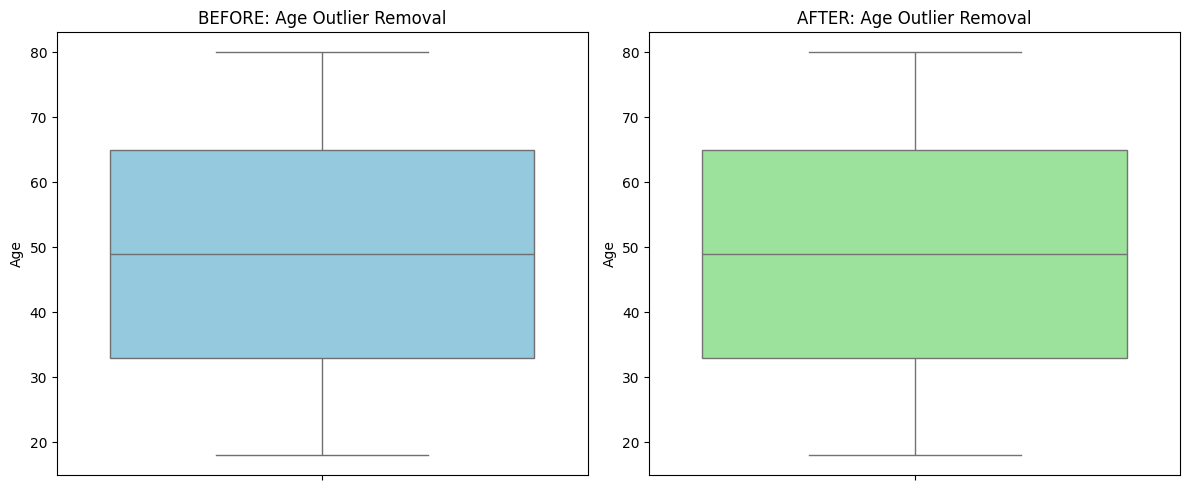
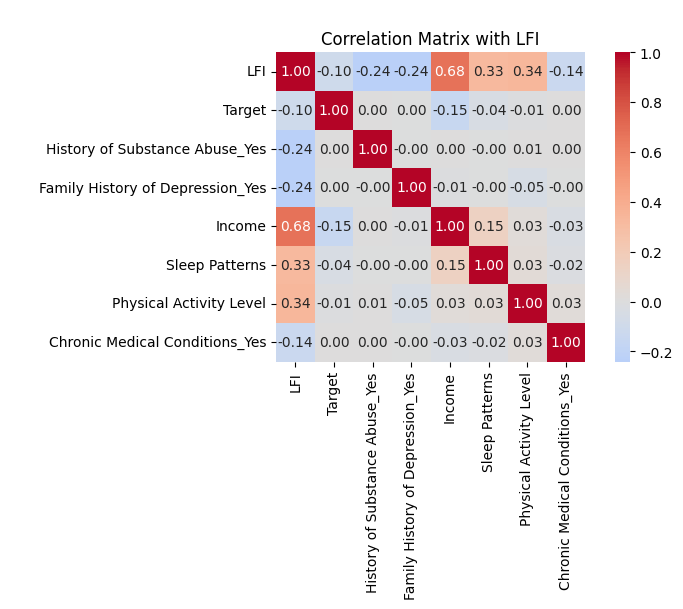
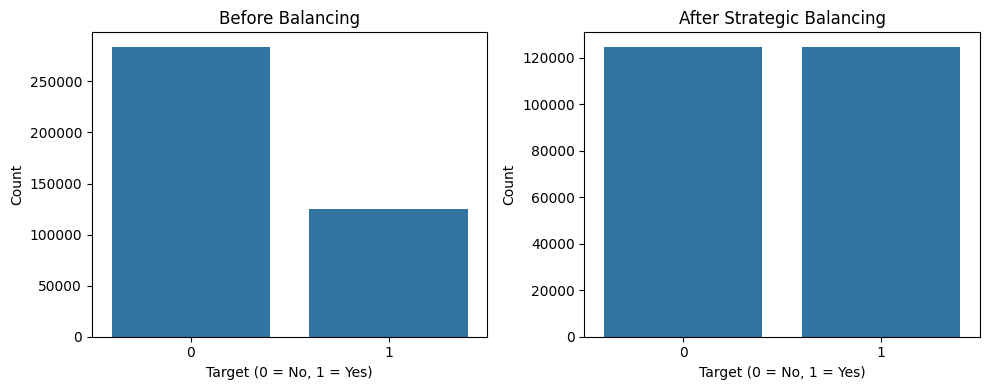
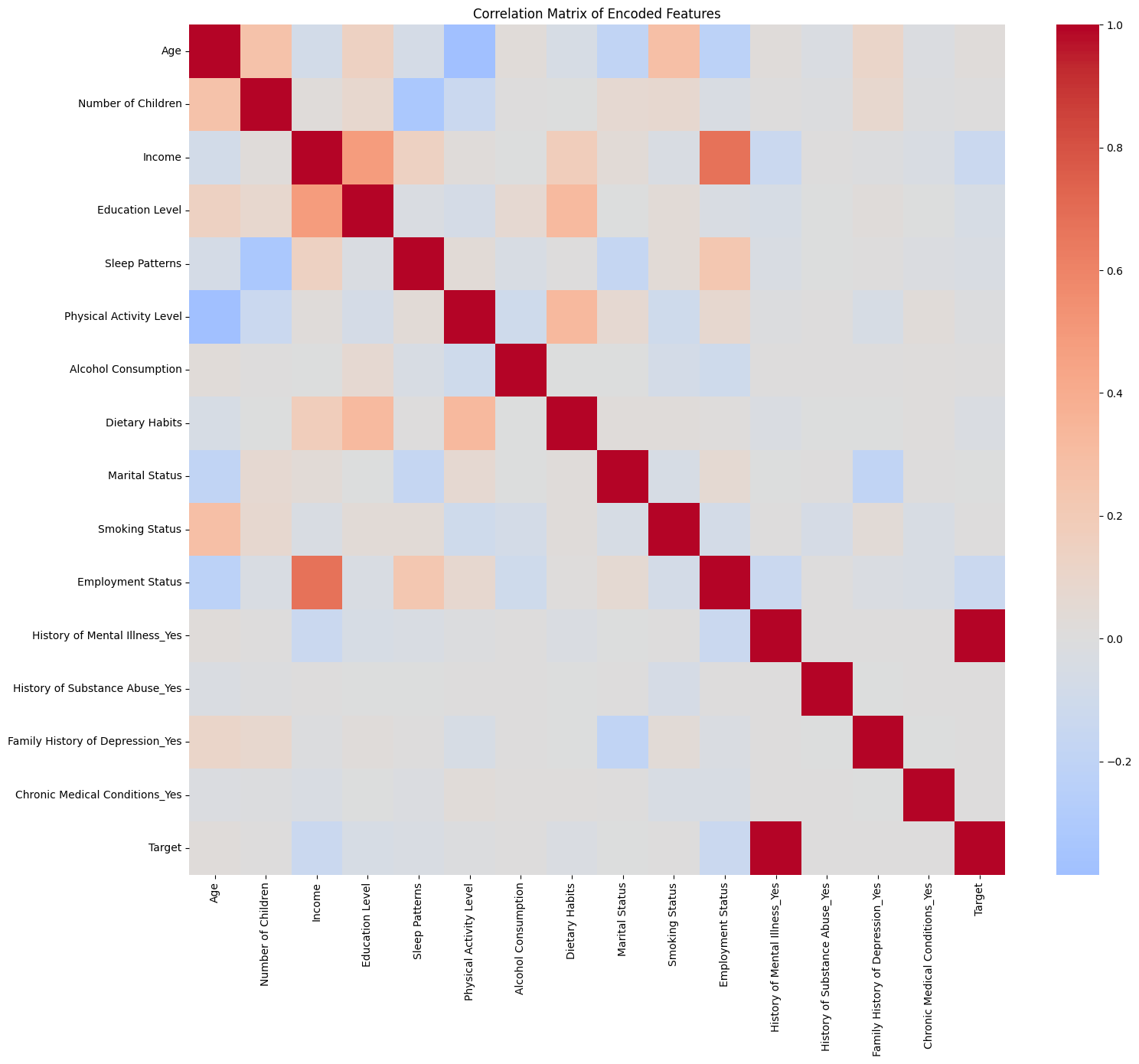
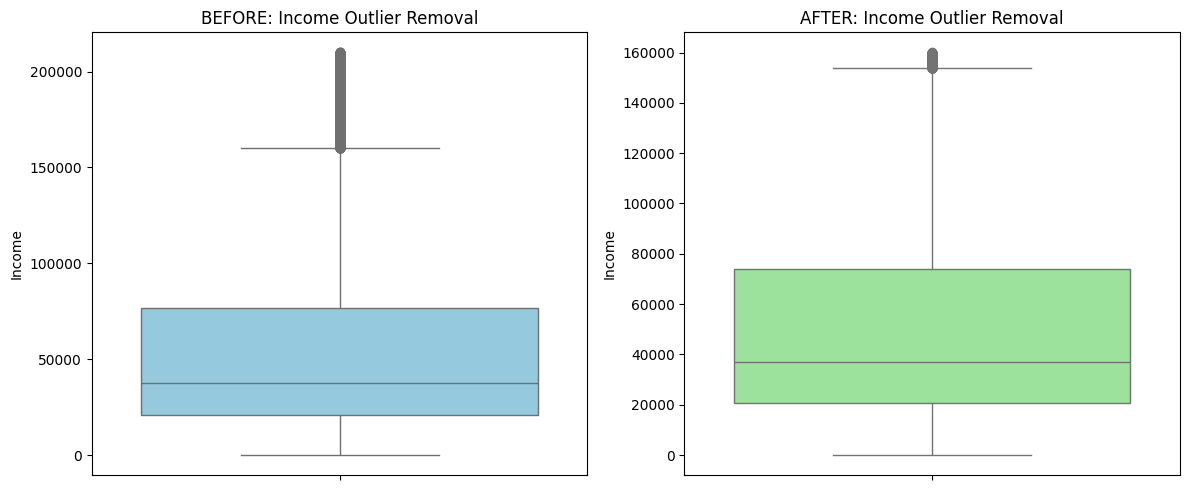
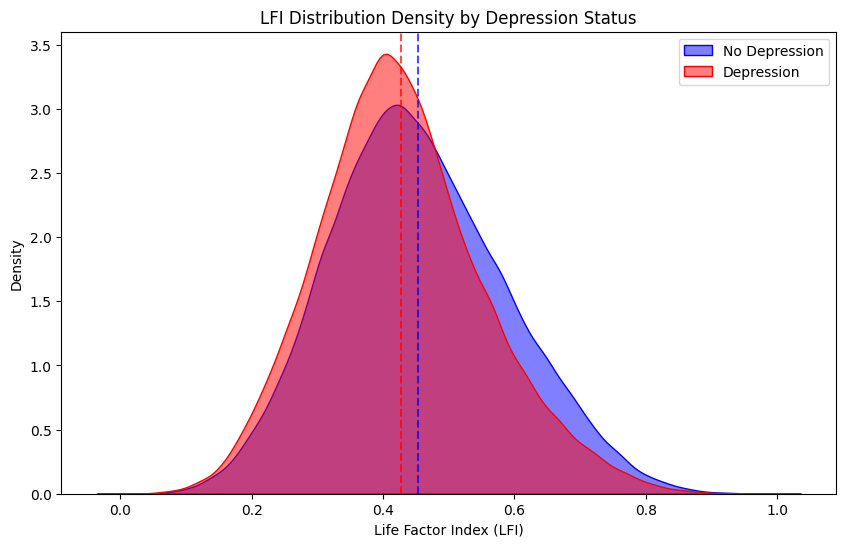
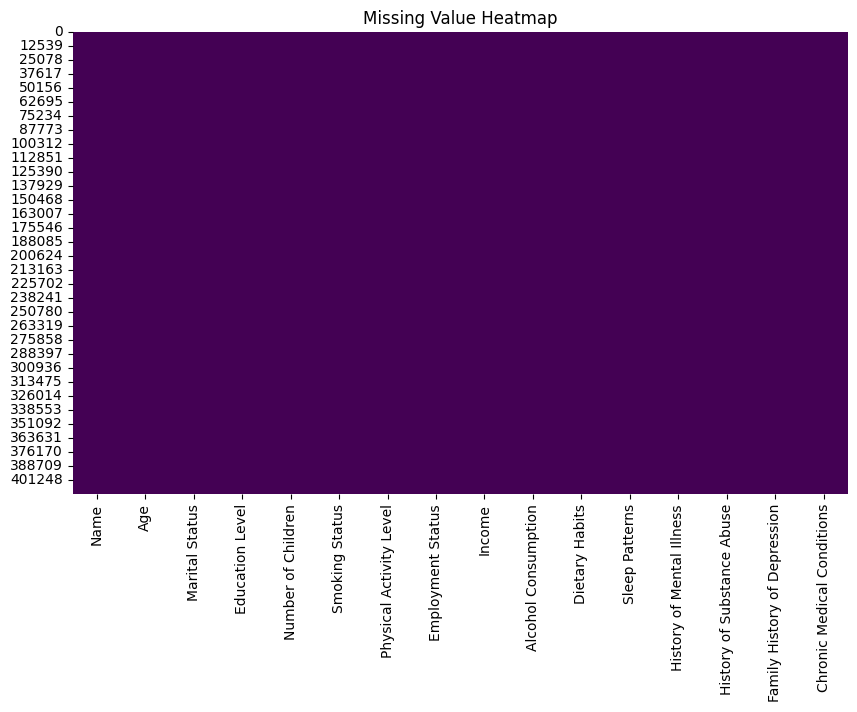
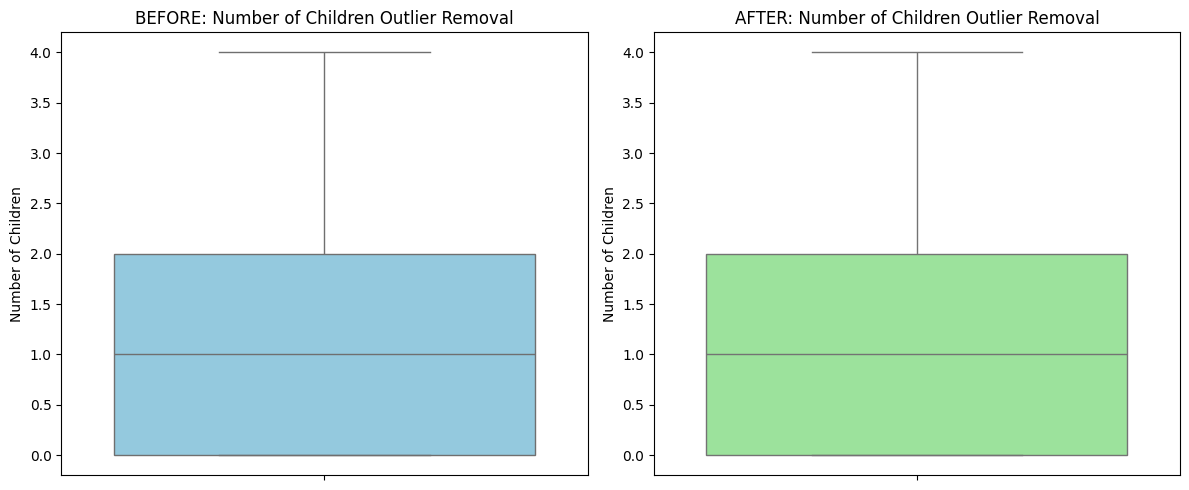
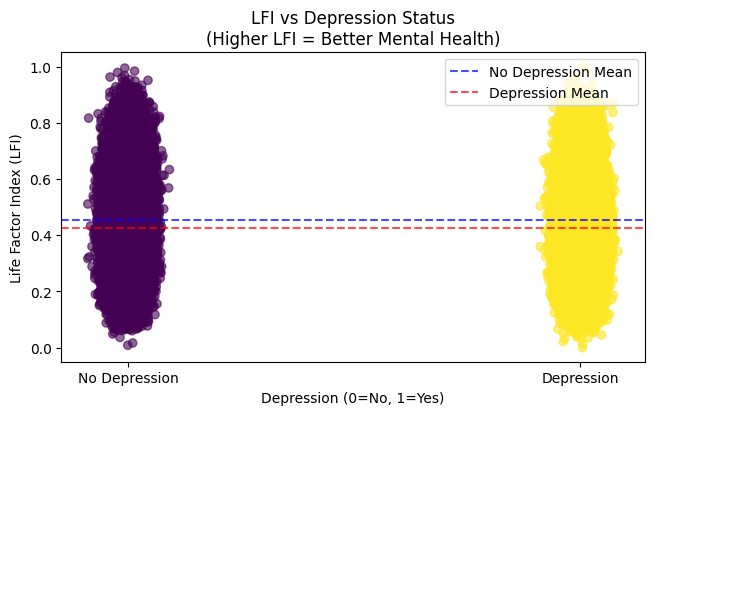
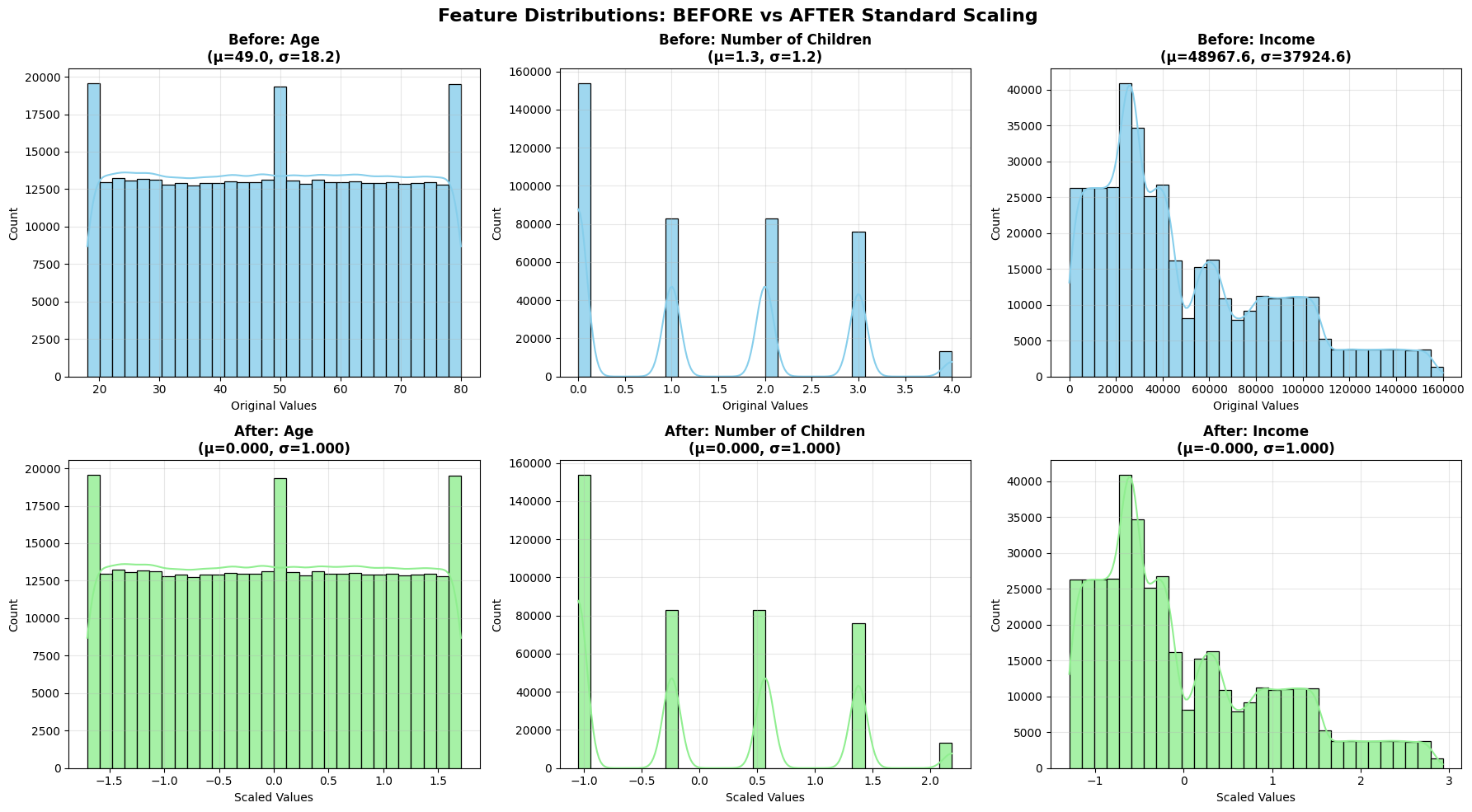
* Created a weighted mental health index using key features like Income, Sleep Patterns, Substance Abuse History, etc.
* LFI combined individual contributions and interactions, scaled to 0–1.
* Used LFI to explore patterns and correlations with depression.

3.6 Data Splitting

* Split the data into training and testing sets for the model training

**3.6 Exploratory Data Analysis**

* Visualized distributions, correlations, and LFI patterns.
* Identified top contributing factors to depression through feature importance and correlation analysis. Insights guided feature selection and model design.



**4. Model Design and Implementation**

Model Selection

Several machine learning algorithms were considered:

* Logistic Regression: A baseline linear model.
* Random Forest: An ensemble method that handles non-linear relationships.
* XGBoost: A gradient boosting algorithm known for its performance.
* CatBoost: A gradient boosting algorithm optimized for categorical features.
* MLP (Multi-layer Perceptron): A neural network model.
* LightGBM: A gradient boosting framework that uses tree-based learning algorithms.

Model Implementation

Each model was implemented in google colab using scikit-learn, XGBoost, CatBoost, and LightGBM libraries. Hyperparameters were tuned using GridSearchCV with cross-validation to optimize performance.

**5. Evaluation and Comparison**

The models were evaluated using the following metrics:

* Accuracy: Percentage of correctly classified instances.
* F1-Score: Harmonic mean of precision and recall.
* ROC-AUC: Area under the Receiver Operating Characteristic curve.

| Model | Accuracy | F1-Score | ROC-AUC |
| --- | --- | --- | --- |
| Logistic Regression | 0.573 | 0.546 | 0.593 |
| Random Forest | 0.572 | 0.589 | 0.594 |
| XGBoost | 0.576 | 0.510 | 0.582 |
| CatBoost | 0.576 | 0.601 | 0.601 |
| MLP | 0.577 | 0.541 | 0.602 |
| LightGBM | 0.572 | 0.588 | 0.595 |

Best Model: CatBoost achieved the highest F1-Score and ROC-AUC, indicating its superior performance.

**6. Ethical Considerations and Bias Mitigation**

**Potential Biases**

* **Demographic Bias:** Overrepresentation of certain age groups and regions can lead to biased predictions.
* **Sampling Bias:** The dataset may not capture the full spectrum of individuals at risk of depression.

**Mitigation Strategies**

* **Balanced Sampling:** Ensure equal representation of different demographics during model training.
* **Bias Detection:** Implement fairness-aware algorithms to detect and mitigate biases.
* **Transparency:** Use explainable AI techniques to make model decisions interpretable.

**7. Reflections and Lessons Learned**

Throughout this project, we encountered challenges related to data preprocessing, model selection, and evaluation. Key lessons include:

* **Importance of Data Quality:** Clean and representative data is crucial for building effective models.
* **Model Interpretability:** Understanding model decisions is essential, especially in sensitive areas like mental health.
* **Ethical Responsibility:** Developers must be aware of and address potential biases in AI systems.

**8. References**

* Therrien, A. (2021). Depression Dataset. Kaggle. Retrieved from <https://www.kaggle.com/datasets/anthonytherrien/depression-dataset/data>
* Github Repository For the project(https://github.com/IT24102476/Life-Factor-Index-AIML-Y2S1)