Predicting Mental Illness: A Data-Driven Approach

1. Key Insights from the Data

- **Class Imbalance**: The dataset is imbalanced, with 69.6% negative cases (No mental illness) and 30.4% positive cases (Yes mental illness).

- **Important Predictors**: The top three predictors of mental illness are Income, Age, and Number of Children.

- **Health Factors**: Chronic medical conditions, substance abuse history, and family history of depression are significant predictors.

- **Lifestyle Factors**: Alcohol consumption, physical activity, and sleep patterns also play important roles.

2. Model Performance and Prediction

- **Logistic Regression**:
  - Accuracy: 69.45%
  - Poor performance on positive cases (0% recall for 'Yes' class)
  - Likely affected by class imbalance
- **Random Forest**:
  - Accuracy: 65.77%
  - Better balanced performance (88% recall for 'No', 15% for 'Yes')
  - More suitable for this imbalanced dataset

3. Model Assessment and Potential Biases

- **Model Choice**: Random Forest outperforms Logistic Regression for this task.

- **Performance**: Moderate overall accuracy, but struggles with predicting positive cases.

- **Potential Biases**:
  1. Class Imbalance Bias: Models favor majority class prediction.
  2. Feature Selection Bias: Some important factors might be missing.
  3. Socioeconomic Bias: Heavy reliance on income as a predictor.
  4. Age Bias: Strong influence of age in predictions.

1. Address class imbalance (e.g., oversampling, SMOTE).
2. Consider ensemble methods or advanced techniques like XGBoost.
3. Collect more data on positive cases if possible.
4. Investigate interactions between features.

5. Include more diverse socioeconomic indicators.

6. Explore non-linear relationships, especially with age.