Springboard-DSC Program

Capstone Project 3

Mental Health and Well-being: Predicting the Need for Mental Health Support

By Jude M. Santos

October, 2024

Table of Contents

1. Introduction

Mental health issues are becoming increasingly prevalent across the United States, yet access to timely, personalized care remains a significant barrier. The lack of adequate support not only harms individuals but also creates far-reaching legal and financial repercussions. Healthcare providers, such as Kaiser Permanente, the Priory Group, Acadia Healthcare, and Universal Health Services, have faced lawsuits, fines, and settlements due to their failure to deliver proper mental health care. For example, Universal Health Services paid $122 million after failing to meet behavioral health standards. These cases highlight the considerable risks and financial consequences of neglecting mental health care, underscoring the urgent need for proactive, comprehensive interventions.

There is a pressing business need for an effective and scalable solution that predicts and supports mental health needs. By leveraging data-driven insights, early intervention can be ensured, helping to mitigate both personal and organizational risks while fostering better mental health outcomes.

By utilizing machine learning models trained on extensive health datasets, this project aims to provide predictive insights into mental health needs, enabling early intervention and resource allocation.

Research Questions

- How can machine learning models help predict mental health support needs?
- Which demographic and behavioral features most strongly influence mental health predictions?
- What is the optimal tradeoff between precision, recall, and cost-benefit in classification models?

Objectives

- Utilize CDC health datasets to build predictive models.
- Clean, preprocess, and analyze data to extract meaningful insights.
- Implement feature engineering to enhance prediction accuracy.
- Train and evaluate multiple machine learning models.

- Optimize models through hyperparameter tuning and cost-benefit threshold adjustments.
- Deploy the final model using Google Cloud (GCP), Docker, and Kubernetes.

Stakeholder Impact

This initiative holds significant relevance for various stakeholders.

- For healthcare providers, it offers tools to prioritize care and allocate resources effectively,
- Mental health professionals benefit from enhanced predictive insights to personalize treatments.
- Policymakers can leverage aggregate data to shape community-level interventions, and
- Individuals gain timely access to mental health resources and support.

The analytical findings, including model performance metrics and insights, will guide strategic interventions in mental health support.

Implementation and Project Artifacts

The following GitHub repositories contain all project files, including data processing, model development, and deployment artifacts:

Data Processing & Model Development

- Data Acquisition & Wrangling
- Exploratory Data Analysis (EDA)
- Feature Engineering & Modeling

Deployment & Infrastructure

- User Interface for Model Interaction
- Google Cloud Platform (GCP) Infrastructure

2. Approach

Data Acquisition and Wrangling

The dataset was sourced from **CDC Behavioral Risk Factor Surveillance System (BRFSS)**, containing responses from the years **2022 and 2023**. The final dataset, after merging and cleaning, contained **700K+ rows and 70 columns**.

Column descriptions are provided in the accompanying codebook files, while The original XPT were data files downloaded from CDC:

- USCODE22_LLCP_102523.HTML (CDC Link)
- USCODE23_LLCP_091024.HTML (CDC Link)
- LLCP2022.XPT (1.16GB) (CDC Link)
- LLCP2023.XPT (1.2GB) (CDC Link)

After the initial acquisition, the data underwent several data-wrangling processes:

- The 2022 and 2023 datasets were converted from the original XPT format to csv format, and kept a file copy of the csv file.
- The datasets were merged, leaving only columns that are common to both datasets. The merge produced +700K rows with 70 columns.
- **Target Feature**: The target variable categorizes mental health status into 4 classes

Class	Description
Class 0	0 days of poor mental health (Healthy group)
Class 1	1–13 days of poor mental health (Moderate group)
Class 2	14+ days of poor mental health (Severe group)
Class 3	Unsure/missing information (Low importance)

- **Handling Categorical and Numeric Mixed Features**: Columns such as **EXEROFT1** and **PHYSHLTH** were cleaned by grouping or mapping mixed values into meaningful categories:

  EXEROFT1

Original Values	Mapped/Grouped Values
Weekly	Weekly
Monthly	Monthly
777	Uncertain
999	Refused
777 (Missing)	Missing

  PHYSHLTH

Original Values	Mapped Values
1–30	Retained as numeric values
88	0 (None)
77	-1 (Don't know/Not sure)
99	-2 (Refused)
BLANK	-3 (Missing)

- **Addressing Missing Values (BLANK)**: For columns labeled as BLANK (Not asked or Missing), these were recategorized to **0 (None)** where applicable.
- **QSTLANG Categorization**: Reclassified values **3–99** into a single category **3 (Other)**.
- **Landline vs. Cell Phone Survey Data:** Combined redundant columns into new unified columns, then Dropped the original columns after combining.

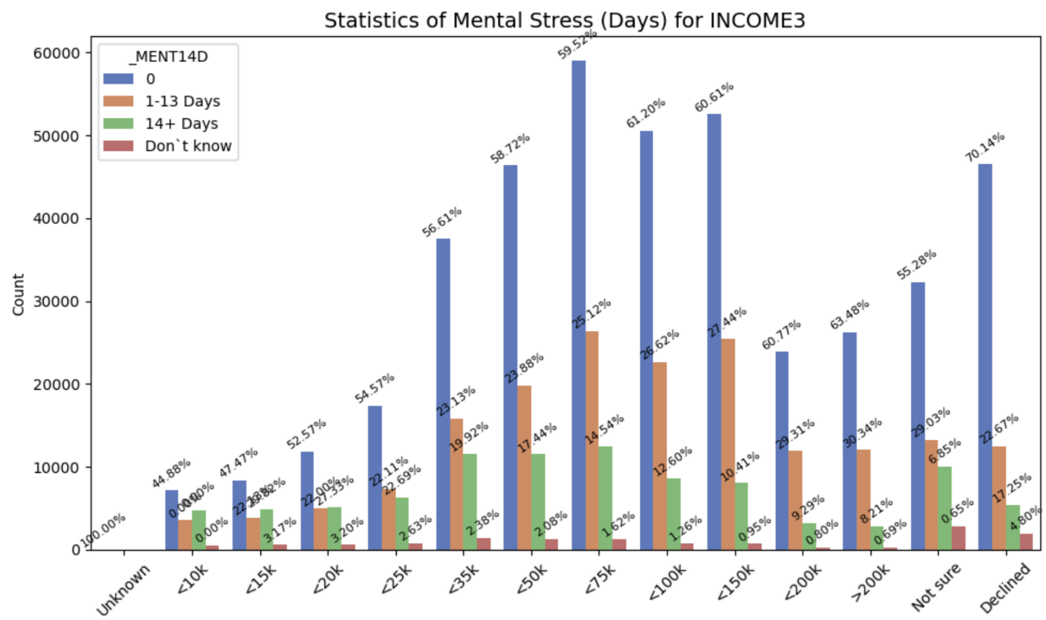
New Feature	Combined Original Features
SEX	SEXVAR, LANDSEX2, CELLSEX2, LNDSXBRT, CELSXBRT, BIRTHSEX
_STATE	_STATE, STATERE1, CSTATE1
ADULT	LADULT1, CADULT1
_METSTAT	_URBSTAT, _METSTAT

- The final column count after combining landline and cellphone data is at **59** columns.
- Converted all features to Integer or Category types

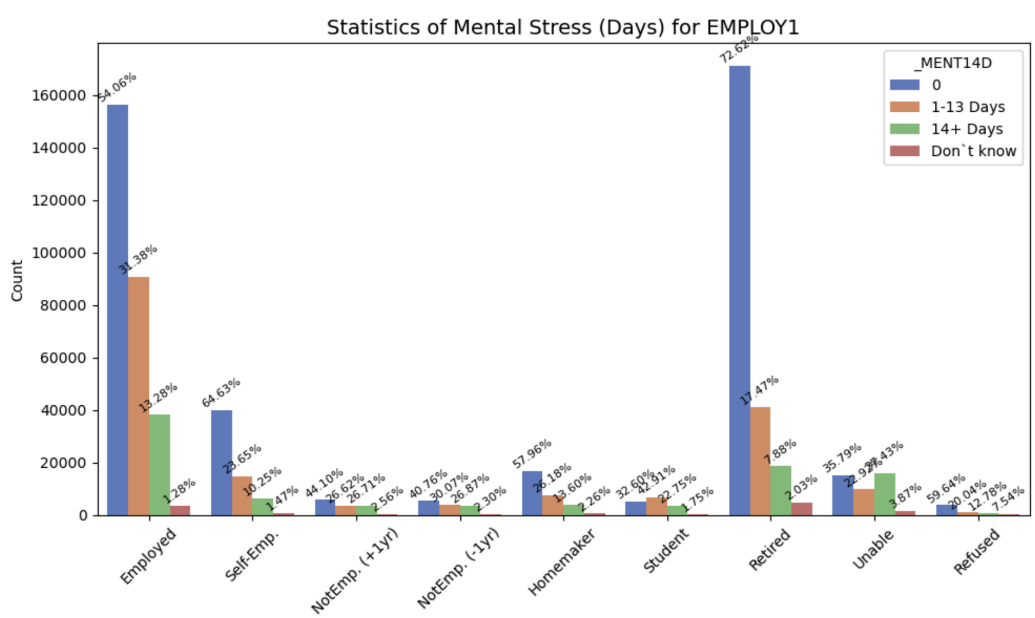
These cleaning steps standardized and streamlined the dataset, ensuring consistent formats and meaningful categories for further analysis.

Storytelling and Inferential Statistics

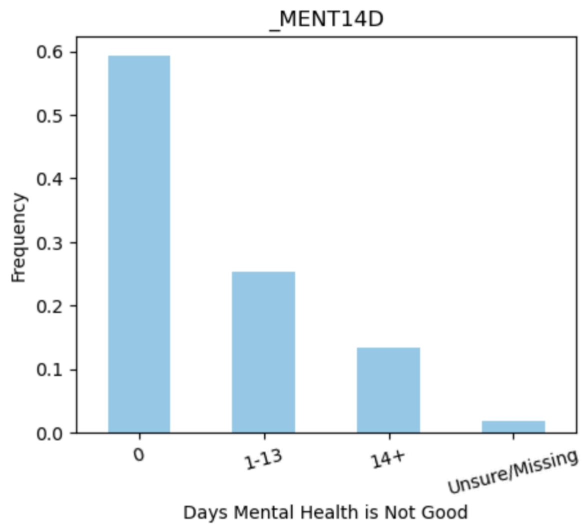
The analysis showed demographic disparities in mental health outcomes, with **stress-free** days dominating most groups, but **stress levels increasing** in **lower-income** brackets, **unemployed individuals**, and certain demographics. **Higher income** earners and **employed** groups show a larger proportion of **stress-free** individuals,



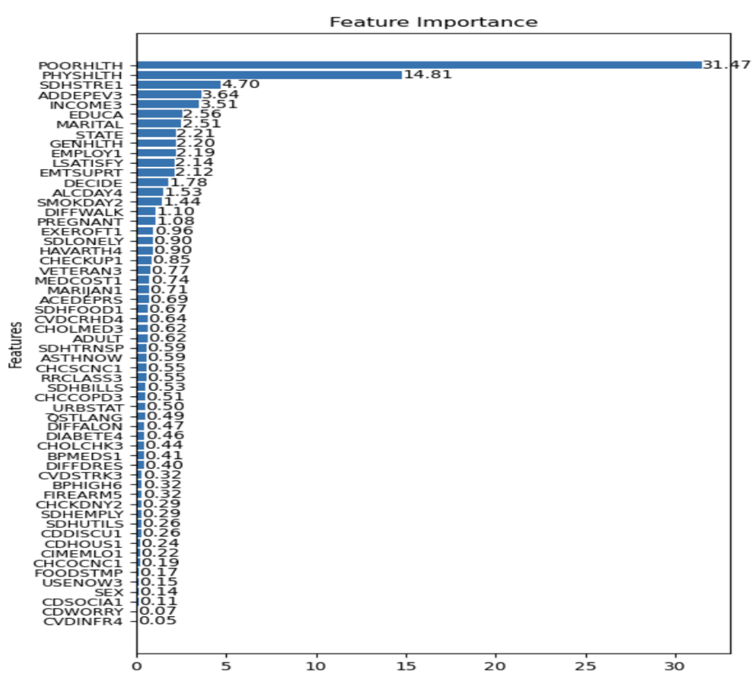
while **economic and social hardships** correlate with **increased stress**.



The target variable (**_MENT14D**) exhibits a significant class imbalance, with most observations falling into the **"0 days"** category. This imbalance suggests the need for strategies to ensure fair representation across all classes, particularly underrepresented groups, to avoid skewed predictions.



Key features, such as **POORHLTH** and **Mental_Health_Composite**, came out as strong predictors of mental health status, while features like **PHYSHLTH and ADDEPEV3** also contributed moderately. Exploring interactions among socioeconomic, behavioral, and health indicators may reveal additional insights. Features with low predictive importance were identified for potential removal to simplify the model without sacrificing accuracy.



The results emphasize the importance of addressing class imbalance through techniques such as resampling (e.g., SMOTE) and the need for **stratified sampling** to maintain representative distributions. Advanced metrics like **precision, recall, and F1-score** are critical to evaluating model performance comprehensively.

Baseline Modeling

The baseline model implemented **Logistic Regression**, which serves as a standard approach for binary classification. This initial model set a benchmark against more complex models. Logistic Regression allowed us to gain insights into feature importance, aiding in the understanding of the factors that contributed significantly to predicting mental health needs.

Model	Accuracy	Macro Avg (P/R/F1)	Weighted Avg (P/R/F1)
Logistic Regression	63%	0.49 / 0.57 / 0.49	0.71 / 0.63 / 0.66

The model was evaluated using **precision** and **recall** metrics, which are crucial for understanding the balance between identifying individuals in need of support and minimizing false positives. Achieving **optimal precision and recall scores** in this base model will serve as a baseline for cost trade-off analysis.

Extended Modeling

To enhance prediction performance and accuracy, we explored more advanced models, including **XGBoost**, **LightGBM**, and **Random Forest**. The motivation for these models stemmed from their ability to handle complex relationships and interactions within the data. The implementation of these models also included hyperparameter tuning, which further optimized their performance.

3. Findings

Model	Accuracy	Macro Avg (P/R/F1)	Weighted Avg (P/R/F1)
Random Forest	80%	0.72 / 0.55 / 0.56	0.79 / 0.80 / 0.79
LightGBM	76%	0.58 / 0.69 / 0.60	0.81 / 0.76 / 0.78
XGBoost	77%	0.58 / 0.68 / 0.60	0.81 / 0.77 / 0.78

The **Extended Models**—Random Forest, LightGBM, XGBoost showed significant improvements over the baseline **Logistic Regression** model in accuracy, precision, recall, and F1-score.

Best Overall Performance:

- **XGBoost** achieved the **highest accuracy (77%)**, closely followed by **Random Forest (80%)** and **LightGBM (76%)**.
- XGBoost and LightGBM had **balanced precision, recall, and F1 scores**, making them strong choices for generalization.

Class Imbalance Handling:

- **Random Forest** had a higher **Macro Average Precision (0.72)** but lower recall (0.55), meaning it was better at precise classifications but struggled with recall.

Balanced Performance Across Metrics:

- **LightGBM and XGBoost had the best balance across Precision, Recall, and F1-score**, making them ideal choices for an optimal tradeoff.
- **Weighted averages** for LightGBM and XGBoost (around **0.81 / 0.76 / 0.78**) suggest they generalize well across all target classes.

The best performing model

XGBoost is the **top choice** for overall performance, balancing accuracy, precision, and recall. **LightGBM** is a **strong alternative**, offering similar performance with slightly better recall. **Random Forest**, while achieving the highest accuracy, may struggle with recall in imbalanced classes.

4. Cost Benefit Threshold Optimization

The analysis aimed to identify the optimal threshold that strikes a balance between minimizing costs and maximizing net benefits. The findings indicate that **0.30** is the best threshold for minimizing costs and maximizing benefits. **These results will be applied in the production model to optimize decision-making and resource allocation.**

Key Findings:

- **Optimal Threshold (Min Cost):** 0.30, with a minimum cost of **18,395,800 units**.
- **Optimal Threshold (Max Net Benefit):** 0.30, with a net benefit of **20,644,470 units**.
- **Cost-Benefit Net Difference: 2,248,670 units**, showing that at this threshold, benefits significantly exceed costs.
- **Class-wise Thresholding Analysis:** The best threshold is **0.1 for all classes**, meaning a uniform strategy may work effectively.

Global Cost-Benefit Threshold Results

Threshold	Net Benefit	Total Cost	TP	FP	FN
0.1	11,466,370	22,356,500	111,269	30,056	30,056
0.2	18,519,990	19,065,750	113,590	27,735	27,735
0.3 (Optimal)	**20,644,470**	**18,395,800**	112,228	29,097	29,097
0.4	19,653,250	18,984,300	110,563	30,762	30,762
0.5	14,611,080	21,236,350	108,008	33,317	33,317

- Best Threshold: **0.3**, where **net benefit is maximized** and **total cost is minimized**.
- Beyond 0.3**:** Costs start increasing, and net benefit decreases.

Class-Wise Cost-Benefit Threshold Results

Class	Optimal Threshold	Precision	Recall	Net Benefit	Total Cost	Net Diff	TP	TN	FP	FN
0 Days (No Stress)	0.1	0.847	0.968	14,961,500	1,274,900	13,686,600	81,182	42,778	14,654	2,711
1-13 Days (Moderate)	0.1	0.484	0.968	13,104,700	4,269,300	8,835,400	34,748	68,408	37,038	1,131
14+ Days (Severe)	0.1	0.306	0.962	5,139,500	13,083,500	-7,944,000	18,223	81,154	41,235	713
Unsure	0.1	0.051	0.885	23,170	0	23,170	2,317	95,278	43,430	300

- Total Net Difference: 14,601,170 units, indicating a strong positive outcome.
- Best Threshold for All Classes: 0.1, meaning uniform thresholding across classes is effective.

Takeaways:

- The threshold of **0.30** is optimal for overall cost-benefit tradeoff.
- Class-wise optimization recommends 0.1 for all classes, maintaining high recall and balancing cost-effectiveness.
- Threshold adjustments do not significantly impact validation and test performance, confirming model stability.

Cost Threshold Matrix Basis

The cost and benefit values assigned to false positives (FP), false negatives (FN), and true positives (TP) were designed as relative and measurable values, meant to provide a structured approach for optimizing decision thresholds. However, these values do not directly correspond to actual medical or financial costs. Instead, they were chosen to represent the relative importance of different types of misclassifications in the context of mental health predictions.

The guiding principles for these assignments were:

- False Negatives (**FN**) incur the highest cost because failing to identify individuals with severe mental health conditions could lead to a lack of necessary intervention.
- False Positives (**FP**) have moderate costs since an incorrect classification may lead to unnecessary resource allocation but is less critical than missing true cases.
- True Positives (**TP**) are assigned high benefits to reflect the positive impact of correctly identifying those in need of support.
- Class 3 (**Unsure**) has minimal costs since misclassifying uncertain cases has the least practical consequence.

Cost-Benefit Assignment Table

Class	False Positive (FP) Cost	False Negative (FN) Cost	True Positive (TP) Benefit
0 Days (No Mental Stress)	50	200	200
1-13 Days (Moderate Stress)	100	500	500
14+ Days (Severe Stress)	300	1,000	1,000
Unsure	0	0	10

Notes:

- These values are not linked to actual healthcare costs but serve as relative weights to guide the threshold selection process.
- The purpose is to find a threshold that maximizes net benefit while minimizing unnecessary costs in a way that aligns with the model's objectives.

These cost assumptions were used in the Cost-Benefit Analysis to optimize thresholds, ensuring that the model effectively prioritizes cases based on mental health severity while balancing overall resource efficiency.

5. Summary

XGBoost came out as the **best model** for **deployment**, with better gains in accuracy, recall, and overall metrics performance.

- Features were reduced to the most relevant subset based on feature importance scores and a minimum importance score of 0.1 or greater.
- Composite feature engineering added meaningful predictors, boosting the model's interpretability and predictive accuracy.
- Training with the refined dataset optimized the balance between overfitting and generalization.
- Threshold adjustments using cost-benefit analysis fine-tuned predictions.

The introduction of **composite features** and **cost-benefit thresholding** proved crucial to achieving these improvements. While the model excels in "0 Days" and "1-13 Days" predictions, further work is recommended for refining predictions in the "14+ Days" and "Unsure" classes.

Cost-Benefit Analysis

- **Global**
    - Optimal Thresholds: Tested at various points (0.1 to 0.5).
    - Net Benefit Peak: Achieved at a threshold of 0.3, with a Net Benefit of 70.
    - Trade-offs: As thresholds increase: True Positives (TP): Peak at a threshold of 0.3 with 340 TP, then decline. False Positives (FP): Gradually decrease as thresholds increase, reaching a minimum at 0.3. False Negatives (FN): Minimized around 0.3 but increase at higher thresholds.
    - The net difference is 2,248,670 units, indicating a positive outcome, as the total benefits exceed the total cost.
- **Class-Wise**
    - The best threshold is 0.1 for all classes, indicating a uniform threshold strategy may work for this dataset.
    - The recall values are relatively on the higher range for the first 3 classes, Class 3 ("14+ Days")is slightly lower.
    - Net Benefit and Total Cost vary widely among classes, reflecting the differing importance or impact of predictions in each class.
    - The net difference is **14,601,170** units, indicating a positive outcome, as the total benefits exceed the total cost.

6. Recommendations

The final model could be considered fit for limited deployment focusing on the "0 Days" and "1-13 Days" classes, provided that the following recommendations are implemented to mitigate the identified limitations.

1. **Human-in-the-Loop Validation**
    - Model predictions for Classes 1, 2, and 3 should flag cases for expert review rather than being final decisions.
    - A team of clinicians or mental health experts should review flagged predictions to validate and prioritize interventions.
    - This minimizes risks associated with false positives, especially for Classes 2 and 3.
2. **Intervention Plan and Resource Allocation**
    - Design an intervention plan commensurate to the confidence and importance of the prediction(s).
    - High-confidence predictions for Class 2 with >0.7 probability: Allocate maximum resources (therapy outreach, immediate support).
    - Medium-confidence predictions for Class 1 with 0.5-0.7 probability: Allocate moderate resources (surveys, preliminary counseling).
    - Low-confidence predictions for Class 3: Delay action and collect additional data for improved prediction.
    - This ensures limited resources are used efficiently while maintaining operational credibility.
3. **Feedback Loop**
    - Continuously monitor and improve model performance post-deployment.
    - Collect data from actual interventions to identify patterns in false positives and false negatives.
    - Retrain the model periodically using updated data that reflects real-world trends.
    - Set up monitoring for data drift to detect if the input data distribution changes over time.
    - Mental health data is dynamic, and ensuring the model adapts to changing patterns is critical.

- Make it clear to stakeholders that the model supports, but does not replace, clinical judgment.
- Develop and provide user guidelines explaining how to interpret the predictions and the potential risks (false positives).
- Label predictions with confidence levels ("High confidence prediction for Class 1: 0.75").

Defined Deployment Plan

1. **Deploy in a Staged Environment**
   - Start with pilot deployment
   - Focus on specific classes (Classes 0 and 1) in controlled settings, such as a single clinic or region.
   - Gradually expand deployment to include all classes as confidence in predictions improves.

2. **Monitoring and Reporting**
   - Set up real-time dashboards to monitor precision, recall, false positives, and false negatives by class.

3. **Develop Automated Alerts for Critical Predictions**
   - For Classes 2 and 3 (where misclassifications have high consequences), trigger alerts for immediate review.

     For example: Alert is sent to a counselor if the model predicts "14+ days" with a confidence score above 0.6.

4. **Integrate with Existing Workflows**

   Ensure the model integrates seamlessly into existing mental health systems such as:

   - patient records
   - counseling scheduling systems

5. **Feedback loops for updating the model with outcomes.**

- Update model with outcomes
- Record intervention outcomes based on predictions (how many flagged individuals were correctly identified as needing support).

6. **Ethical and Regulatory Considerations**

   In mental health prediction, ethical considerations should be high priority:

   - Avoiding Harm: Ensure the system does not stigmatize individuals by overestimating mental health issues (Classes 2 and 3 false positives).
   - Data Privacy: Comply with regulations like HIPAA to protect sensitive patient data.
   - Bias Mitigation: Conduct fairness analysis to ensure the model performs equitably across demographics (gender, ethnicity).

6. **Evaluation Metrics**

   To evaluate the effectiveness of the contingencies:

   - Precision improvement for Classes 1, 2, and 3 after threshold adjustment.
   - Reduction in false positives for Classes 2 and 3 post-deployment.
   - User feedback from clinicians or mental health experts on flagged cases.
   - Resource utilization rates aligned with predicted severity levels.

The above plan safeguards for the limitations of the model while allowing the system to provide actionable insights.

7. Conclusion

Deploying the mental health classification model is valuable if it is complemented by a detailed workflow plan such as mental-health experts validation, threshold adjustment, and continuous feedback.

For a mental health application, the stakes are high, and ensuring the system complements expert judgment is critical. With these strategies, we can deploy the model responsibly while iterating to improve its reliability.

8. Consulted Resources

- CDC BRFSS Dataset: 2022 Data | 2023 Data
- Python Libraries Used: pandas, scikit-learn, xgboost, lightgbm, seaborn, matplotlib, bayesopt
- Mental Health Trends**:** Our World in Data