Wednesday, October 15, 2025

Predicting Poor Mental Health Days Using Behavioral and Demographic Data from the Behavioral Risk Factor Surveillance System (BRFSS), 2022–2023

Thesis Paper for I-492 Project

by

Abdullah Siddiqui

Advisor
Dr. Sridhar Ramachandran

Table of Contents

I. Abstract

Purpose: Summarize the key aspects of your research.
Content: Include the research hypothesis, methods, key results, and conclusions.
Length: Keep it between 150-250 words.
Tips: Write this section last to ensure it accurately reflects the content of your paper.

Your paper should be formatted single spaced, using Times New Roman Font Size 12.

This study explores how behavioral and demographic factors predict poor mental health days using data from the Behavioral Risk Factor Surveillance System (BRFSS) 2022–2023. The hypothesis is that lifestyle and socioeconomic variables such as sleep, exercise, smoking, alcohol use, income, and education can be used to accurately predict mental health outcomes. Using R and Python, linear and logistic regression models were developed and tested. An interaction term between education and income was included to capture joint effects of socioeconomic status. The logistic regression model achieved an AUC of 0.731, indicating moderate predictive performance. Results showed that lower income, less education, and lack of exercise were associated with higher frequencies of poor mental health days. Findings align with previous studies showing strong behavioral determinants of mental health, while highlighting the importance of interpretable, survey-aware machine learning approaches for public health. These insights can inform targeted interventions and future research using advanced explainable AI methods.

II. Introduction

Background: Provide context and background information on your research topic.
Problem Statement: Clearly state the research problem or question.
Hypothesis: Present your hypothesis or the main objective of your study.
Significance: Explain the importance and potential impact of your research.

Mental health issues represent a growing public health concern, affecting millions of individuals globally and imposing substantial economic and social burdens. Identifying behavioral and demographic risk factors that contribute to poor mental health can help guide prevention strategies. The Behavioral Risk Factor Surveillance System (BRFSS) provides an extensive, nationally representative dataset containing self-reported information on health behaviors and outcomes. This project seeks to use BRFSS 2022–2023 data to predict poor mental health days using statistical and machine learning techniques. The central hypothesis is that behavioral and lifestyle factors, especially exercise frequency and sleep, combined with socioeconomic indicators like education and income, can accurately predict mental health outcomes.

The objectives are threefold: (1) to analyze correlations between behavioral and demographic variables and poor mental health days, (2) to construct and evaluate regression-based predictive models, and (3) to interpret significant predictors to derive practical public health insights.

III. Literature Review

You have already put in a lot of work for this section in your Assignment # 2. Bring that work over here and integrate it with the paper.

Scope: Summarize key findings from existing research related to your topic.
Analysis: Critically analyze the literature, highlighting gaps or inconsistencies.
Relevance: Relate the literature to your research question and hypothesis.

Recent studies have used BRFSS data to investigate behavioral and mental health relationships. The CDC (2022, 2023) confirmed the dataset's validity and representativeness. Ahmadi-Montecalvo et al. (2025) demonstrated the link between social needs and stress among veterans using BRFSS 2022. Similarly, Okeke et al. (2024) and Kamal et al. (2023) found that behavioral variables like sleep, physical activity, and substance use are consistent predictors of mental distress. Kamal (2023) further showed that machine learning can achieve high predictive accuracy when applied to BRFSS data. However, methodological gaps remain, particularly around the lack of survey weighting and limited interpretability. This project builds on that literature by integrating traditional regression with interpretable methods to enhance validity and applicability.

IV. Methods

Dataset: Describe the dataset used, including its source and any preprocessing steps.
Tools: List the tools and software used for analysis.
Procedures: Detail the methods and techniques employed in your research.
Reproducibility: Ensure that another researcher could replicate your study based on this section.

Dataset: Data were drawn from the CDC BRFSS 2022–2023 survey, which includes over 400,000 respondents across the United States. A subset focusing on mental health, behavioral, and socioeconomic variables was created, including poor mental health days, exercise, smoking, alcohol use, education, income, age group, and sex.

Data Preparation: The dataset was cleaned in R by recoding invalid responses (e.g., 77 = 'Don't know', 99 = 'Refused') and removing missing values. The final dataset was saved as BRFSS2023_subset_clean.csv.

Analysis: Descriptive statistics were computed to summarize population characteristics. Correlation analyses and visualizations (scatterplots, heatmaps, cross-tabulations) explored bivariate relationships. Baseline linear and logistic regression models were developed, followed by models including interaction terms (education × income). Model performance was evaluated using ROC/AUC metrics, confusion matrices, and diagnostic plots. All analyses were performed in R, with visualization via ggplot2 and model evaluation using the caret and pROC packages.

V. Results

Presentation: Present your findings clearly and concisely.
Visuals: Include tables, graphs, and charts to illustrate your results.

Narrative: Provide a narrative that explains the significance of the data presented.

Descriptive Statistics: The average number of poor mental health days was 4.36. Approximately 75% of respondents reported regular exercise, 21% smoked, and 47% did not drink alcohol. More than 70% had at least some college education, and over half reported annual incomes above $35,000.

Regression Results: Linear and logistic regression analyses identified significant predictors of poor mental health days. Lower education and income levels were associated with more poor mental health days. Exercise frequency showed a strong protective effect, while smoking and alcohol use were risk factors. Incorporating an education × income interaction revealed that higher education mitigated the adverse mental health effects of low income. The logistic regression model achieved an AUC of 0.731, indicating moderate predictive accuracy.

Visualizations: The GitHub repository contains correlation heatmaps, scatterplots, coefficient plots, and ROC curves illustrating model results (https://github.com/abdullahs1357/SeniorThesis).

VI. Discussion

Interpretation: Interpret your results and discuss their implications.
Comparison: Compare your findings with those from the literature review.
Hypothesis: Discuss whether your results support or refute your hypothesis.
Limitations: Acknowledge any limitations of your study and suggest areas for improvement.

The findings align with previous research showing that behavioral and socioeconomic variables strongly influence mental health outcomes. The predictive model confirms that exercise, education, and income play critical roles in explaining variance in poor mental health days. The observed AUC of 0.731 suggests that even with simple regression models, behavioral and demographic data can moderately predict mental health outcomes. The inclusion of the education × income interaction underscores the complexity of socioeconomic determinants, indicating that higher education may buffer the negative mental health effects of lower income levels.

These results reinforce the need for interpretable and survey-weighted predictive methods when working with population health data. While advanced machine learning models could enhance accuracy, interpretability remains essential for policy application.

Limitations include the reliance on self-reported data, potential residual confounding, and the cross-sectional design of BRFSS, which limits causal inference. Future studies should explore temporal patterns and apply interpretable ML methods such as SHAP or LIME.

VII. Conclusion

Summary: Summarize the main findings of your research.
Implications: Discuss the broader implications of your results.

This study demonstrates that behavioral and demographic factors can moderately predict poor mental health days using BRFSS 2022–2023 data. Regression models highlight that

socioeconomic status and lifestyle behaviors such as exercise significantly influence mental health outcomes. Findings suggest that public health interventions promoting physical activity and educational access could reduce poor mental health prevalence. Future work will apply explainable machine learning to improve interpretability while maintaining predictive performance.

VIII. Future Directions

This project demonstrated that behavioral and demographic variables can moderately predict poor mental health days using BRFSS 2022–2023 data. Future research can extend these findings in several ways:

1. Apply Interpretable Machine Learning (IML) Models:Implement survey-weighted and interpretable ML techniques (e.g., SHAP, LIME, explainable boosting machines) to better understand how individual features influence predictions and improve transparency.
2. Use Longitudinal Data:Combine multiple years of BRFSS data (e.g., 2015–2023) or link to other datasets to examine temporal trends and potential causal relationships.
3. Expand Behavioral Predictors:Include variables related to sleep quality, diet, and access to care to develop a more holistic model of mental health.
4. Public Health Policy Simulation:Test how interventions, such as increased exercise participation or improved access to education, could influence predicted mental health outcomes at the population level.
5. Cross-Dataset Validation:Validate the models using other large health surveys (e.g., NHANES, NSDUH) to confirm the robustness and generalizability of the predictors identified.

List of References

Centers for Disease Control and Prevention. (2022). 2022 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2022.html

Centers for Disease Control and Prevention. (2023). 2023 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2023.html

Ahmadi-Montecalvo, H., et al. (2025). Examining health-related social needs and their association with stress and mental health among US Veterans using 2022 BRFSS data. Discover Public Health, 22(353).

Okeke, M., et al. (2024). Short sleep duration and frequent mental distress: BRFSS 2022. Journal of Behavioral Health, 33(2), 112–120.

Kamal, S., Alharbi, F., & Kumar, M. (2023). Using machine learning to predict poor mental health from BRFSS. The American Journal of Geriatric Psychiatry.

Salvi, S., Roy, A., & Kalagnanam, J. (2025). Classifying tooth loss in the United States using BRFSS 2022 and explainable AI. Electronics, 14(17), 3559.

Pearce, M., et al. (2022). Association Between Physical Activity and Risk of Depression: A Systematic Review and Meta-analysis. JAMA Psychiatry, 79(6), 550–559.

Link to the References

Google Drive Reference Folder: https://docs.google.com/document/d/1FReT1GlA9o9T3MTGsNI4uckBdxV0Ebi0Dq43JGOQDso/edit?usp=sharing

GitHub Repository: https://github.com/abdullahs1357/SeniorThesis

Make sure to store all the sources (article, papers, data set etc.) in a Google Drive Folder or in a OneDrive Folder since you will have to regularly share them with me this semester. For this assignment, include the link to your Google Drive folder or One Drive folder here. Make sure you have added my email (sriramac@iu.edu) to the folder so I can have access to it.