Matthew Maslow
Professor Cleary
CDS-DS-596
May 5th, 2025

Predicting Depression Risk from NHANES Data: Logistic Regression, XGBoost, and SMOTE

Introduction

Depression is one of the many significant public health concern in the U.S., often going undiagnosed despite its broad impact on quality of life and productivity. In 2021, it was recorded that 21 million adults in the U.S. experienced at least one major depressive episode, underscoring the widespread nature of this condition (NIMH, 2023). Standard screening tools like the PHQ-9 provide clinically accepted benchmarks but rely on subjective self-reports and lack integration with broader behavioral or clinical health data. This creates a gap between symptom-based screening and objective data-driven early detection.

This project explores the potential of supervised machine learning to identify patterns in individual-level health data that detect depressive symptoms. Using NHANES 2021–2023, a rich and representative dataset, two PHQ-9 items (DPQ020 and DPQ060) are used as binary targets to model self-reported depressive symptoms. Behavioral, demographic, and biomarker variables from domains like sleep, physical activity, smoking, alcohol use, and inflammation are evaluated jointly to explore how they interact in identifying depression risk.

Recent studies demonstrate the promise of machine learning in mental health, with one broad scoping review synthesizing 300 papers across areas like diagnosis, treatment support, and public health monitoring. Most applications focused on detecting depression and schizophrenia, using algorithms like decision trees, support vector machines, and neural networks. However, the review noted ongoing challenges around generalizability and limited use of standardized, population-representative datasets (Shatte). In contrast, another study applied K-Nearest Neighbors, Decision Trees, and Naive Bayes machine learning methods to a small, survey-based dataset developed during the COVID-19 pandemic using the Hamilton Depression Rating Scale. Designed with psychiatric input, the custom survey aimed to identify individuals at risk of depression during a period of heightened emotional distress, suggesting that machine learning could supplement conventional screening tools in crisis scenarios (Sofia).

My project differs in both scale and scope. Rather than focusing on crisis-specific or clinical data, it applies interpretable machine learning methods, Logistic Regression, and XGBoost to a large, nationally representative dataset: NHANES 2021–2023. It specifically models two binarized PHQ-9 indicators (DPQ020 and DPQ060) as targets for depressive symptoms and incorporates a broad set of predictors spanning behavioral, demographic, and clinical biomarker domains. Additionally, this project prioritizes transparency and accessibility using SHAP for model interpretation and SMOTE to address class imbalance, contributing to more robust, generalizable tools for early depression detection in public health contexts.

Objectives

This project aims to evaluate how well behavioral, socioeconomic, and clinical health factors from NHANES can predict the risk of depression on self-reported symptoms. By modeling two target indicators, DPQ020 (feeling down/hopeless) and DPQ060 (feeling bad about oneself), the study compares the performance of interpretable models (Logistic Regression) and ensemble learners (XGBoost). The secondary goal is to identify which variables contribute most significantly to depression risk predictions and evaluate whether class imbalance handling, like SMOTE, improves model performance.

Data Description

This study uses publicly available data from the 2021–2023 National Health and Nutrition Examination Survey (NHANES), a biannual program administered by the CDC that captures demographic, behavioral, and clinical information from a nationally representative sample of U.S. residents. Two binary target variables were derived from the Patient Health Questionnaire (PHQ-9) depression screener items:

- *DPQ020_binary*: Encodes whether the respondent reported feeling down, depressed, or hopeless in the past two weeks.
- *DPQ060_binary*: Encodes whether the respondent reported feeling bad about themselves, like a failure, or letting their family down.

Each variable was binarized such that “Not at all” = 0, and all other depressive responses (“Several days,” “More than half the days,” or “Nearly every day”) were coded as 1.
Predictor variables span several NHANES modules:

- *Demographics*: Age (RIDAGEYR), gender (RIAGENDR), race/ethnicity (RIDRETH3), marital status (DMDMARTZ), education (DMDEDUC2), income-to-poverty ratio (INDFMPIR)
- *Income*: Monthly poverty index (INDFMMPI, INDFMMPC), family savings >$5,000 (INQ300)
- *Sleep*: Weekday and weekend sleep hours (SLD012, SLD013)
- *Smoking & Alcohol*: Smoked 100+ cigarettes in life (SMQ020), ever drank (ALQ111), past-year drinking frequency (ALQ121)
- *Physical Activity:* Moderate activity minutes (PAD800), sedentary minutes per day (PAD680)
- *Clinical Biomarkers:* Glycohemoglobin (LBXGH), C-reactive protein (LBXHSCRP), and vitamin D (LBXVIDMS)

Responses coded as 7, 9, 77, 99, 7777, and 9999, which represent refused, do not know, or otherwise invalid entries, were removed across all variables prior to analysis, as they contributed noise and no meaningful information for modeling. All numeric predictors were standardized using StandardScaler. An 80/20 stratified train-test split was performed to ensure class balance across outcome variables. The final dataset integrates diverse health behaviors, socioeconomic indicators, and clinical markers to enable interpretable machine-learning approaches for predicting depression risk.

Methodology

Following data cleaning and preprocessing, two binary classification models were developed for each depression indicator (DPQ020_binary and DPQ060_binary): Logistic Regression and XGBoost. All predictor variables were standardized using StandardScaler after removing noisy or non-informative codes (e.g., 7, 9, 77, 99, 9999). An 80/20 stratified train-test split ensured that class proportions were preserved in both training and test sets.

Logistic Regression was implemented as a baseline model within a SMOTE pipeline to test oversampling effects. The regularized logistic model (L2 penalty) was tuned using validation curves across inverse regularization values (C). XGBoost was trained using manually selected parameters (learning_rate=0.05, n_estimators=1000, max_depth=10) based on performance trends from validation curves. The class imbalance was addressed in two ways: using scale_pos_weight in default models and using SMOTE as a separate oversampling strategy on the training data.

Model performance was assessed using several approaches. Five-fold cross-validation was applied to the training set to estimate F1-score with mean and standard deviation. Validation curves were used to visualize model accuracy trends across hyperparameters, C for Logistic Regression and max_depth for XGBoost. Confusion matrices were generated on the test data to provide insight into false positives and negatives, particularly for minority-class predictions. Final evaluation metrics included accuracy, precision, recall, and F1-score.

Each modeling path was run with and without SMOTE to explore the impact of class imbalance adjustments. This enabled a direct comparison between native class weighting in Logistic Regression and synthetic oversampling for both classifiers. Finally, SHAP summary bar plots were generated for each XGBoost model to identify the top contributing features. These plots supported interpretability and informed the decision to drop variables with low predictive power or redundancy, such as ALQ270 and PAD820.

Results

Model performance was assessed across two binary classification targets, DPQ020_binary (felt down/depressed) and DPQ060_binary (felt bad about oneself), using both Logistic Regression and XGBoost. Each model was tested in two configurations: class weighting and SMOTE oversampling to evaluate how it handles imbalanced target distributions.

For DPQ020_binary, class-weighted Logistic Regression reached a test accuracy of 62%. Recall for the depressed class (1) was 0.69, indicating moderate sensitivity, while precision for that class was 0.45. The non-depressed class (0) achieved higher precision (0.79) and recall (0.59), showing that the model favored the correct identification of non-depressed cases. XGBoost scored slightly higher on overall accuracy (65%), with stronger performance on non-depressed individuals (precision: 0.72, recall: 0.80), but recall for the depressed class dropped to 0.38, suggesting reduced sensitivity despite better precision balance.

Introducing SMOTE helped recover some sensitivity to the depressed class at the cost of overall precision. SMOTE-enhanced Logistic Regression maintained similar recall (0.66) while slightly lowering depressed-class precision. For XGBoost with SMOTE, depressed-class recall marginally improved to 0.32, but the model showed signs of weaker class separation and overfitting tendencies.

For DPQ060_binary, the class-weighted Logistic Regression model achieved a higher test accuracy of 66%, with balanced recall (0.65) and precision (0.41) for the depressed class. Non-depressed performance remained strong (precision: 0.84, recall: 0.66). XGBoost again posted a higher accuracy (69%), but depressed-class recall fell to 0.27, showing a familiar trade-off between total accuracy and minority-class sensitivity. SMOTE preserved Logistic Regression's recall while having little impact on XGBoost performance for this target.
5-fold cross-validation confirmed that Logistic Regression models had more consistent F1-scores: 0.53 (DPQ020) and 0.47 (DPQ060), compared to 0.40 and 0.34 for XGBoost.
SHAP summary plots showed that age (RIDAGEYR) was consistently the top predictor. Other influential features included weekday sleep (SLD012), vitamin D (LBXVIDMS), income-to-poverty ratio (INDFMPIR), and biomarkers like glycohemoglobin (LBXGH) and CRP (LBXHSCRP). While top features shifted slightly between targets, behavioral and clinical variables proved jointly important to prediction.

Discussion & Interpretation

These results suggest that combining diverse NHANES variables, including behavioral, clinical, and socioeconomic indicators, has meaningful predictive value for identifying self-reported depressive symptoms. While overall recall (i.e., sensitivity to actual positive cases) remained moderate, the models consistently flagged variables like age, C-reactive protein (LBXHSCRP), glycohemoglobin (LBXGH), family income ratio, and sleep duration as top predictors. These patterns align with existing knowledge on the biological and social determinants of mental health.

The recall is significant in a mental health screening context because failing to identify individuals experiencing depressive symptoms (false negatives) could delay critical care or intervention. In contrast, false positives are more easily managed through follow-up. For this reason, optimizing sensitivity, even at the expense of precision or accuracy, can be a worthwhile trade-off. Techniques like SMOTE or threshold calibration may further improve performance in this area.

Logistic Regression models with class weighting or SMOTE achieved more balanced and interpretable results across evaluation metrics. XGBoost models delivered higher accuracy but disproportionately predicted the non-depressed class, raising concerns about class imbalance and potential deployment bias if left uncorrected.

Limitations include the reliance on self-reported symptom data (rather than clinician-verified diagnoses), lack of external validation on out-of-sample populations, and the risk of overfitting, particularly in the XGBoost models. Moreover, NHANES offers only cross-sectional snapshots, which prevents the modeling of symptom progression or temporal risk trajectories. Despite these constraints, the findings support the potential for low-cost, population-based screening models built on public health data. Such models may be particularly valuable in digital health platforms or community outreach programs where comprehensive clinical screening is infeasible.

Conclusion

This project demonstrates the feasibility of using interpretable machine learning to predict depressive symptoms from publicly available health survey data. By leveraging NHANES 2021–2023 and modeling two binary indicators from the PHQ-9 screener, the study highlights the predictive value of demographic, behavioral, and clinical variables, particularly age, sleep patterns, income status, and inflammation markers. Logistic Regression consistently offered a balanced trade-off between accuracy and sensitivity, while SHAP analysis provided meaningful insights into feature importance. Although limitations such as self-report bias and lack of external validation remain, the findings suggest a practical path toward scalable, low-cost screening tools that could augment early identification and outreach efforts in mental health care. Future work should explore model generalization, longitudinal data, and mobile or digital health platform integration.

Works Cited

Kroenke, K et al. “The PHQ-9: validity of a brief depression severity measure.” Journal of general internal medicine vol. 16,9 (2001): 606-13. doi:10.1046/j.1525-1497.2001.016009606.x

“Major Depression.” *National Institute of Mental Health*, U.S. Department of Health and Human Services, www.nimh.nih.gov/health/statistics/major-depression. Accessed 25 Apr. 2025.

“Nhanes Questionnaires, Datasets, and Related Documentation.” *Centers for Disease Control and Prevention*, wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023. Accessed 26 Apr. 2025.

Shatte, Adrian B R et al. “Machine learning in mental health: a scoping review of methods and applications.” Psychological medicine vol. 49,9 (2019): 1426-1448. doi:10.1017/S0033291719000151

Sofia, et al. “Machine Learning Based Model for Detecting Depression during Covid-19 Crisis.” *Scientific African*, U.S. National Library of Medicine, 13 May 2023, pmc.ncbi.nlm.nih.gov/articles/PMC10182866/#sec0003.