Title:
From Patterns to Predictions: A Multi-Model Exploration of Depression Risk

Author(s): Alexis Parker, Vikram Vaddamani, Brandon Fox

Introduction

Depression is a major public health concern in the United States, with widespread effects on individual well-being and healthcare systems. This project uses data from the 2021–2023 NHANES Questionnaire to investigate how behavioral, medical, and social factors relate to depression severity. We applied unsupervised learning techniques to identify subgroups in the population that share similar characteristics, and trained nine supervised learning models to predict PHQ-9-based depression severity categories. Three unsupervised methods and nine supervised models were implemented across the team. The workflow included data cleaning, feature engineering across multiple NHANES modules, and derived income estimates using external references. Feature importance was evaluated using SHAP to interpret key predictors of model performance. Compared to related work, this project uses a recent and nationally representative sample, incorporates multi-module survey data, and emphasizes interpretability across modeling phases. Initial findings suggest that sleep, education, chronic illness, insurance status, and income are consistently important in identifying individuals at higher risk of moderate to severe depression.

Related Work

The predictive modeling of depression using machine learning has become an increasingly valuable approach to understanding mental health outcomes, particularly in the wake of the COVID-19 pandemic, which has altered social dynamics, healthcare access, and psychological stressors globally. One foundational study that guided this project is Prediction of depressive disorder using machine learning approaches: findings from the NHANES by Vu et al. (2025). This study used pre-pandemic data from the 2013–2014 National Health and Nutrition Examination Survey (NHANES) to apply a range of supervised machine learning (ML) techniques—Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and LightGBM—to predict moderate to severe depression, operationalized as a PHQ-9 score ≥10. The authors highlighted the interpretability of models using SHAP values to assess feature importance and found high predictive utility across several models, demonstrating the potential of ML to outperform traditional diagnostic tools. Our project builds upon this work by incorporating some the same supervised learning techniques (Logistic Regression, Random Forest, SVM) and interpretability framework while introducing a novel combination of unsupervised learning methods to generate new features that enhanced supervised prediction. Importantly, we utilized post-COVID NHANES data, which reflects a vastly changed mental health landscape and thus extends the relevance and timeliness of depression modeling efforts.

Complementing Vu et al.'s ML-driven analysis is the study by Cheong Kim (2023), Exploring Factors Influencing Depression: Socioeconomic Perspectives Using Machine Learning Analytics, which analyzed data from the Korean National Health and Nutrition Examination Survey (KNHANES). This study emphasized the socioeconomic determinants of depression, including variables like income, marital status, and working hours, revealing cultural and economic nuances unique to the South Korean context. Notably, it found that higher income was sometimes associated with greater depression, challenging typical assumptions and emphasizing stress-related mechanisms. While Kim's work used logistic regression without expanding into more complex ML methods, its exploration of SES factors informed our own variable selection process and motivated a deeper investigation into post-pandemic socioeconomic shifts in the U.S. population. Our work advances this by incorporating a broader array of unsupervised learning techniques, capturing non-linear relationships between socioeconomic factors and depression that traditional regression may overlook.

A third relevant study is the Importance of Social Determinants in Screening for Depression by Califf et al. (2022), which used cross-sectional data from the Baseline Health Study (BHS) to explore how PHQ-9 scores correlate with social determinants of health. This research emphasized the significance of demographic and contextual variables—such as access to care, housing, and income stability—in shaping mental health outcomes, aligning with a growing consensus that effective screening and prediction must extend beyond biological and behavioral metrics. While Califf et al. focused primarily on descriptive analysis within a clinically rich cohort and did not apply ML techniques, their emphasis on the PHQ-9 and social determinants directly supports our approach. We adopted a similarly broad lens but translated these insights into predictive features for both unsupervised and supervised learning, offering an empirically grounded yet scalable ML framework for mental health prediction in the general population.

By integrating supervised learning methods from Vu et al., drawing socioeconomic inference strategies from Kim, and adapting the public health screening context explored by Califf et al., our study offers a unique contribution. Specifically, we expand the modeling of depression into the post-COVID era, deploy unsupervised methods to generate features that reflect latent clusters of risk, and maintain methodological continuity with established ML frameworks—all while updating the analysis to reflect contemporary health trends and social disruptions.

Data Sources

This project uses data from the 2021–2023 cycle of the National Health and Nutrition Examination Survey (NHANES), a national program conducted by the National Center for Health Statistics (NCHS), part of the U.S. Centers for Disease Control and Prevention (CDC). NHANES collects health, nutrition, and demographic information from a representative sample of the U.S. population using in-person interviews and standardized physical exams. The data are publicly available, widely cited in epidemiological research, and considered a high-quality source for studying health outcomes at the population level. For this project, raw questionnaire data files were downloaded from the CDC website in .xpt format and converted to .csv for analysis.

Seven questionnaire modules were included in the final dataset: Demographics (DEMO), Depression Screener (DPQ), Functioning (FNQ), Health Insurance (HIQ), Hospital Utilization and Access to Care (HUQ), Income (INQ), and Sleep Disorders (SLQ). After merging by respondent ID (SEQN), the combined dataset contained nearly 12,000 records and 93 variables. Records were filtered to include only adults aged 18 and older who completed both the interview and physical exam. Placeholder values and invalid response codes were removed, and column names were updated for clarity. The PHQ-9 total score was calculated by summing responses to nine depression screener items. Following standard clinical scoring rules, respondents were only retained if they answered at least six of the nine items. The final cleaned dataset contained 467 rows and 26 features, and was used for both unsupervised and supervised learning tasks.

Feature Engineering

Feature engineering was a critical step in preparing the NHANES dataset for both unsupervised and supervised machine learning modeling of depression outcomes. The initial dataset, composed of multiple NHANES modules, included a wide range of demographic, behavioral, medical, and mental health variables. The cleaning and selection process began with filtering for relevant respondents—adults aged 18 and older—ensuring the sample reflected populations typically assessed for clinical depression using the PHQ-9. Participants with missing PHQ-9 values, or with inconsistent or invalid response codes across any key variables, were removed to ensure model accuracy and validity.

Raw variables were then renamed for clarity and standardization. For example, NHANES variable codes such as “DPQ010” (feeling down, depressed, or hopeless) were renamed to more interpretable labels. This approach improved readability during modeling and aligned with the broader goal of making findings accessible to non-technical stakeholders. Non-numeric response codes (e.g., “Refused,” “Don’t Know”) were either mapped to NaN values and imputed if necessary, or led to row exclusion if they occurred in variables deemed critical.

Next, a selection process was conducted to isolate variables with conceptual relevance to depression. Variables included in the final modeling dataset spanned demographic attributes (age, sex, race/ethnicity, marital status, education level, income-to-poverty ratio), health behaviors (sleep duration, alcohol consumption, smoking status, physical activity), and self-reported general and mental health status. Redundant or clinically irrelevant variables, such as those tied to highly specific lab values or conditional follow-up questions only posed to narrow subgroups, were dropped. Additionally, variables with more than 30% missing data or no meaningful variation across the sample were excluded.

To enhance the socioeconomic context of each NHANES respondent, we constructed a derived feature by joining external data from the U.S. Census Bureau. Rather than using NHANES's limited or occasionally suppressed income data directly, we leveraged publicly available Census statistics to estimate individual-level income based on demographic traits. Specifically, we collected national median income values broken down by education level, race/ethnicity, and gender. For each respondent in the NHANES dataset, we identified their educational attainment, racial identity, and gender, and then matched each of those characteristics to its corresponding Census-based median income. If a respondent had values available for all three traits, we extracted the median income from each subgroup and calculated the average of those three figures to create a personalized estimated income score. This approach allowed us to assign a consistent, demographically-informed income estimate to every respondent, even when actual income data was missing or reported in obscure ranges. The resulting feature, joined to the NHANES data using the SEQN respondent identifier, was used as a socioeconomic predictor in our supervised learning models.

The final dataset used for modeling consisted of a curated set of cleaned and engineered features representing biological, behavioral, and socioeconomic dimensions associated with depression. A complete mapping of original NHANES codes to cleaned variable names, as well as a list of included and excluded features, is provided in the appendix.

Unsupervised Learning

Unsupervised learning techniques were used to investigate patterns in PHQ-9 depression symptom severity among adult NHANES respondents. The overarching goal was to uncover latent subgroups that shared similar characteristics across domains such as health behavior, socioeconomic status, and access to care. To this end, multiple clustering approaches were explored, including K-Means,

PCA-informed visualizations, and DBSCAN. These methods allowed for the examination of structure within the data without relying on predefined labels. Among these approaches, K-Means emerged as the most stable and interpretable when applied to the standardized, preprocessed dataset. The following analysis presents key findings from the K-Means clustering workflow, including cluster composition, dominant feature patterns, and implications for supervised modeling.

K-Means Clustering of NHANES Respondents

All selected variables were numeric at the time of clustering. Nominal variables such as insurance coverage and healthcare access indicators were one-hot encoded. Ordinal and binary variables retained their original numeric encoding, and continuous features like age, poverty index, and sleep hours were preserved. To address missing data, two versions of the dataset were prepared: a drop version that included only complete cases (n = 273), and an imputed version (n = 467) that used median or mode imputation depending on variable type. Both datasets were standardized using StandardScaler before clustering.

Cluster selection was informed by the Elbow Method and Silhouette Score across a range of k values. In the imputed dataset, the silhouette score peaked at k = 4, and the inertia curve showed a visible inflection around the same point. While the drop version exhibited lower overall inertia, its silhouette scores were consistently lower and less stable across values of k. The k = 4 solution was selected for its balance between interpretability and consistency across both data versions.

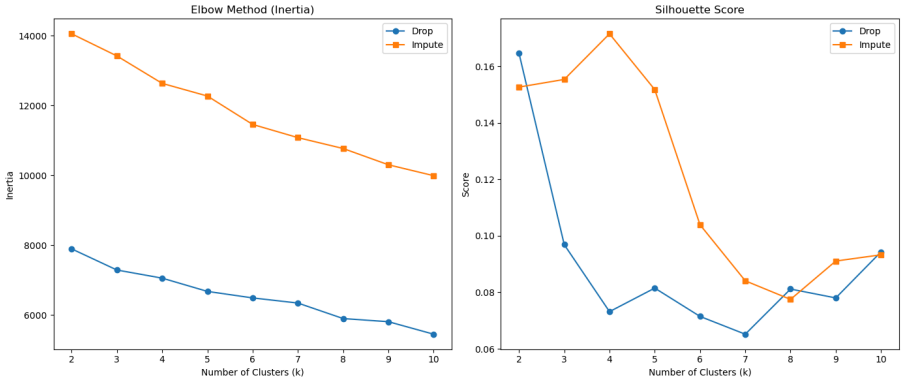


Figure 1. *Elbow and Silhouette plots comparing cluster evaluation metrics across drop and imputed datasets.*

PHQ-9 total scores were examined by cluster to evaluate differences in depression symptom severity. As shown in Figure 2, the distribution of scores varied across clusters in both the drop and imputed versions. Median PHQ-9 scores and distribution shapes differed meaningfully, indicating that the clustering algorithm captured subgroups with distinct symptom levels.

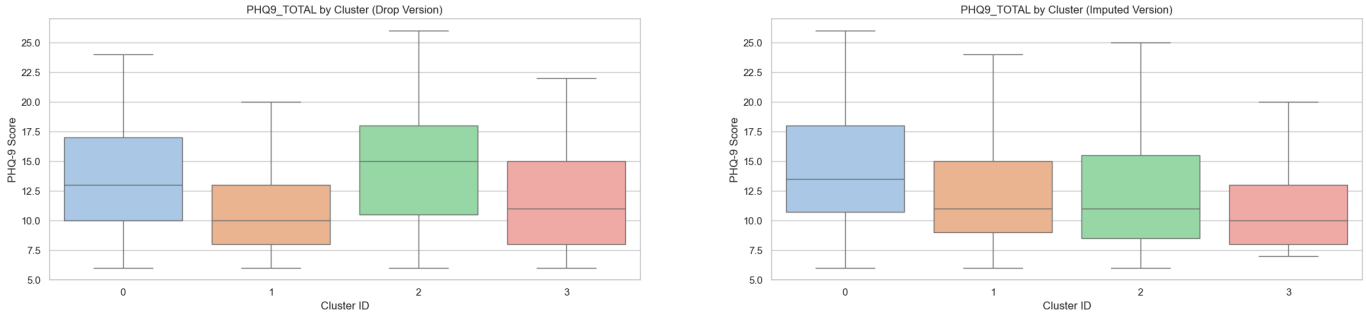
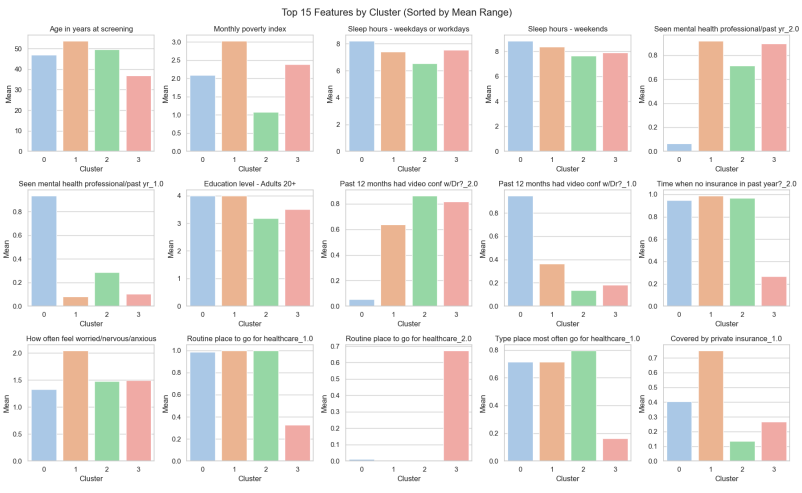


Figure 2. *PHQ-9 total score distributions by cluster for drop (left) and imputed (right) datasets.*



Key features were ranked based on the range of their mean values across clusters. Figure 3 displays the top 15 features from the drop version, including age, weekday sleep hours, poverty index, and use of mental health care. These variables reflect demographic, behavioral, and access-related dimensions that co-occur with differences in depression risk. A full comparison between the drop and imputed versions, including shared and unique features, is provided in Appendix A.

Figure 3. *Top 15 features ranked by mean value range across clusters (drop version).*

K-Means clustering revealed distinct subgroupings based on combinations of social, behavioral, and health-related characteristics. These patterns offer a foundation for interpreting structure in the data and informed the selection of priority variables for downstream supervised learning.

Method: Principal Component Analysis

PCA is one of the dimensionality reduction techniques that we used. PCA excels at showcasing the features that explain the variance in the dataset, allowing the selection of the most influential features. This way of selecting the features with the highest variance is also a noise reduction technique that is helpful in dimensionality reduction.

Data Preparation and Feature Encoding

- The features were encoded in the same method used in K-Means Clustering. This allows all columns to be numerical (which is needed for future model steps)
- The features were imputed using the median of each feature to ensure that there were no missing values
- Standardization was done on all features to normalize the scale.

All of these methods were done as the PCA model requires no missing values and all numerical values in order to run. Standardization was necessary to make the results of the analysis interpretable.

Component Analysis

Before the model was run the parameter, n_components, was set to 0.95. This is done so that the features that explain at least 95% of the variance in the dataset are obtained.

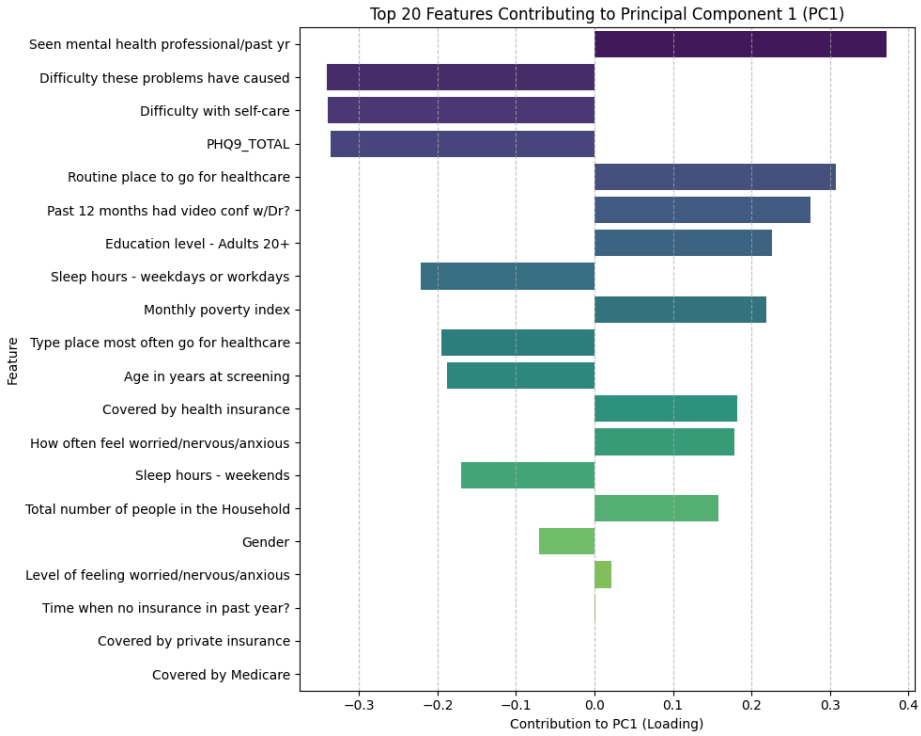


Figure 4: *Top 20 features that were produced from the PCA model analysis. The direction of the bars explain the signed influence that the features have.*

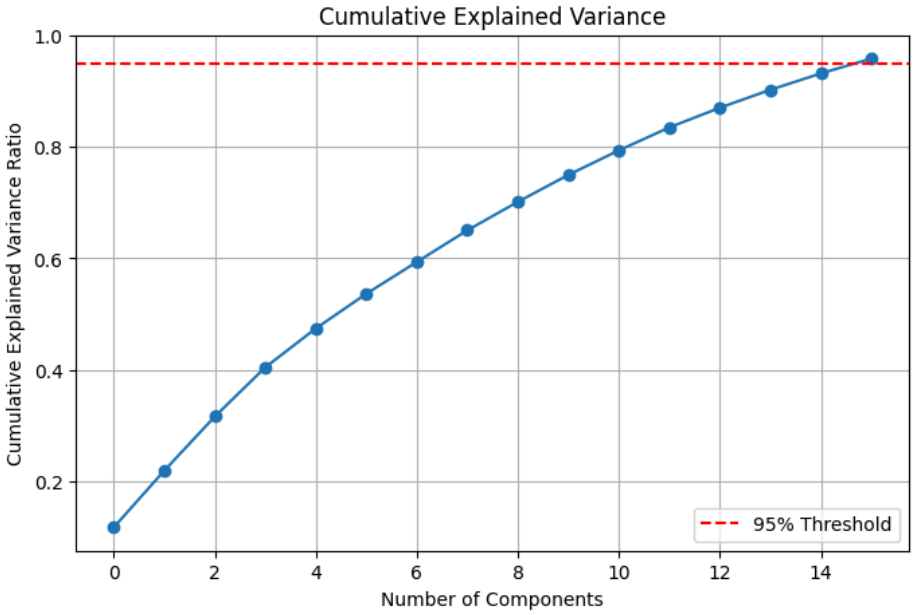


Figure 5: *Shows the cumulative variance that is explained by the features with each additional feature. The logarithmic curve shows that the earlier components explain larger amounts of the variance and it tapers off as the last 5 features are reached.*

Reflection

PCA modeling is an excellent technique at reducing the noise in a dataset performing dimensionality reduction. The fact that the hyperparameters can be tuned so that the features can explain X amount of the variance or a top Y amount of features can be selected allows users to see how influential the features are and get rid of non influential features or noise. This method helped in the performance of the supervised learning.

Top 15 Features Identified Across Clustering and Dimensionality Reduction Analyses

KMeans	**PCA**	**DBSCAN**
Age in years at screening	Seen mental health professional/past yr	Usual sleep time weekdays/workdays
PHQ9_TOTAL	Difficulty these problems have caused	Usual sleep time weekends
Monthly poverty index	Difficulty with self-care	Usual wake time weekends
Sleep hours - weekdays or workdays	PHQ9_TOTAL	Usual wake time weekdays
Sleep hours - weekends	Routine place to go for healthcare	Age in years at screening
Routine place to go for healthcare [There is no place][1]	Past 12 months had video conf w/Dr?	Total number of people in household
Routine place to go for healthcare [yes]	Education level - Adults 20+	Ratio of family income to poverty

[1] **Survey question:** Is there a place that {you/SP} usually {go/goes} when {you are/he/she is} sick or {you/s/he} need{s} advice about {your/his/her} health?

Type place most often go for healthcare [Emergency room][2]	Sleep hours - weekdays or workdays	Level of feeling depressed
Type place most often go for healthcare [A VA medical center or VA outpatient clinic]	Monthly poverty index	Level of feeling worried/nervous/anxious
Type place most often go for healthcare [A doctor's office or health center]	Type place most often go for healthcare	How often feel worried/nervous/anxious
Education level – Adults 20+	Age in years at screening	Education Level - Adults 20+
Covered by military health care	Covered by health insurance	Gender
Past 12 months had video conf w/Dr? [No][3]	How often feel worried/nervous/anxious	Difficulty remembering or concentrating
Past 12 months had video conf w/Dr? [Yes]	Sleep hours - weekends	Past 12 months had video conf w/Dr?
How often feel worried/nervous/anxious	Total number of people in the Household	Seen mental health professional/past yr

Dataset Preparation

Before attempting hyper parameter selection, the dataset was imputed using median imputation and standardized. The data was split into a training set and a test set using the stratify parameter to ensure equal class balance in the splits. Originally 'PHQ-9 Total', was a numerical column, but to make the results more interpretable the column was binned and converted from numerical to ordinal. This also reduces the amount of classes that the models have to classify overall improving performance. 'PHQ-9 Total' now has these values instead of numerical ones: "None/Minimal", "Mild", "Moderate", "Moderately Severe", "Severe".

Hyper Parameter Tuning

GridSearchCV and Stratified K-Fold were the methods that were used to find the best hyper parameters for each of the 3 models. These methods were used to perform tuning because GridSearchCV allowed for the optimal hyperparameters to be used and the cross validation method allowed for an even class balance to happen which was crucial for the dataset to preserve.

The selected parameter grids for each model were designed to balance flexibility, stability, and class imbalance. For Logistic Regression, values of C ranged across magnitudes to explore varying regularization strengths, with l2 penalty and the lbfgs solver chosen for their support of multiclass classification. For Random Forest, tuning focused on controlling depth, splits, and leaf size to reduce overfitting, while testing class weighting helped mitigate imbalance. For SVM, tuning spanned linear and RBF kernels, varying C and gamma to test both simple and non-linear boundaries, alongside optional class weighting to improve Severe case detection.

Modeling Approach

Since unsupervised learning was mainly used as a method to conduct dimensionality reduction for the supervised learning models, each team member took an unsupervised approach and continued their work. For example, one member took the features generated from PCA analysis and ran those features on all 3 supervised learning models.. The models that were selected were Logistic Regression, Random Forest, and Supported Vector Machines (SVMs). The target variable that we were interested in was PHQ-9 Total.

After the hyperparameters were set and the data was prepared, the models were each run with X being the features found in the unsupervised learning method and Y being PHQ-9 total. The models were scored using 3 main methods. These methods include

[2] **Survey question:** {What kind of place is it/ What kind of place {do you/does SP} go to most often} - a doctor's office or health center; an urgent care center or clinic in a drug store or grocery store; an emergency room; a VA Medical Center or VA outpatient clinic; or some other place?

[3] **Survey question:** During the past 12 months, did {you/SP} receive counseling or therapy from a mental health professional such as a psychiatrist, psychologist, psychiatric nurse, or clinical social worker?

accuracy, weighted f1-score and the macro-averaged ROC-AUC. The score for each of the 5 classes was calculated and a total score along with a confusion matrix was created.

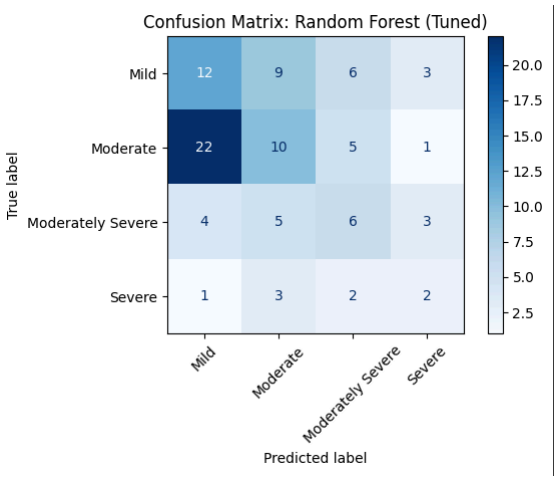


Figure 6: *Example of one of the confusion matrices that were produced from running each model. This graph shows the predicted vs. true labels for the model, showing how well it performed on the data. This particular model was Random Forest which used PCA for its initial feature selection. It scored:*

- **Accuracy**: 0.32
- **Weighted F1-Score**: 0.30
- **ROC-AUC (Macro-Averaged)**: 0.628

Supervised Learning Based on KMeans-Derived Predictors

This analysis used a tailored feature set informed by KMeans clustering results, combined with an engineered socioeconomic variable, Median_Income, estimated from Census tract data and participant demographics. Additional predictors included weekday sleep hours, frequency of anxiety, type of health insurance, and household size. These features were used to train three supervised classifiers: Logistic Regression, Random Forest, and Support Vector Machine.

Stratified five-fold cross-validation with grid search was applied for hyperparameter tuning. Final model performance was evaluated on a held-out test set using accuracy, weighted F1 score, and macro-averaged ROC-AUC, which reflect balanced performance across the four PHQ-9 severity categories.

A fourth model was trained using only socioeconomic variables to assess their predictive strength in isolation (see Appendix B for the confusion matrix). Model performance and misclassification patterns were examined using confusion matrices and SHAP visualizations.

All three models concentrated most predictions in the Mild and Moderate categories, with fewer correct classifications at higher severity levels. Severe cases were especially underdetected. Random Forest demonstrated the most balanced distribution, correctly identifying 2 out of 8 Severe cases and 7 out of 18 Moderately Severe cases. In contrast, both Logistic Regression and SVM failed to correctly classify any Severe cases. Most misclassifications occurred between adjacent categories, reflecting the challenge of differentiating overlapping symptom levels.

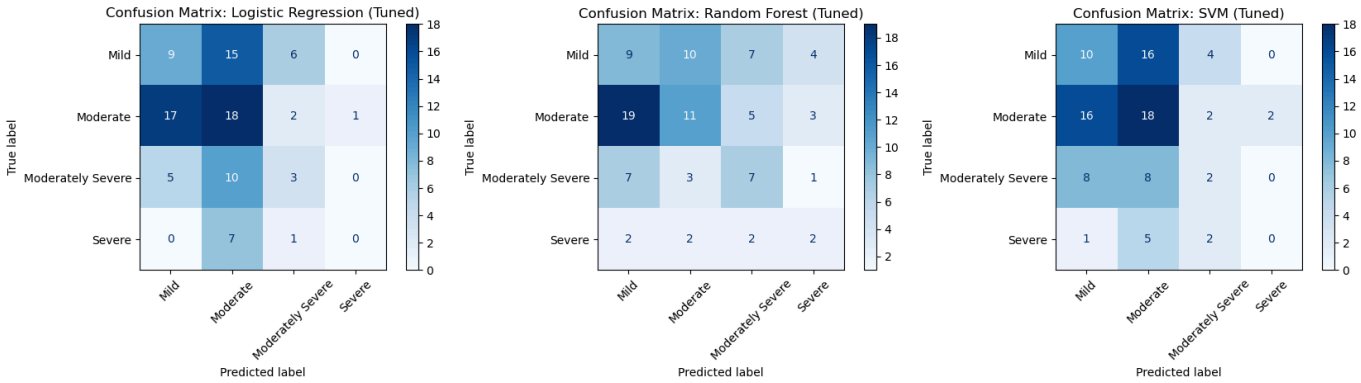


Figure 7. *Confusion matrices for Logistic Regression, Random Forest, and SVM classifiers (tuned).*

Random Forest identified a mix of behavioral and structural predictors as most influential in classifying PHQ-9 severity. Monthly poverty index, age, weekday sleep hours, and weekend sleep hours ranked highest, followed by anxiety frequency. In contrast, insurance indicators such as Medicaid, Medicare, and private insurance coverage contributed less to the model's decisions. This pattern suggests that individual-level behavioral and socioeconomic characteristics played a stronger role than access-related variables in driving predictions.

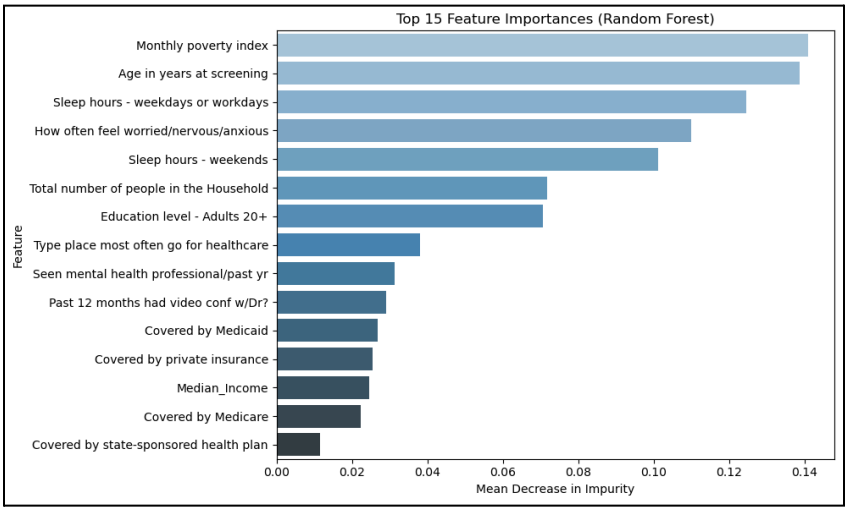


Figure 8. *Top 15 feature importances from the tuned Random Forest model (mean decrease in impurity).*

SHAP values were used to interpret individual predictions and highlight how specific features influenced misclassifications. Figure 8 shows a case where the model predicted *Severe* despite a true label of *Moderate*. The strongest contributor was frequent anxiety, which substantially increased the predicted probability. Additional positive contributors included limited access to care, state-sponsored insurance, and a recent mental health visit. In contrast, living in a larger household and private insurance coverage slightly reduced the predicted risk. This case illustrates how distress-related factors may dominate the model's logic, even when moderating socioeconomic features are present.

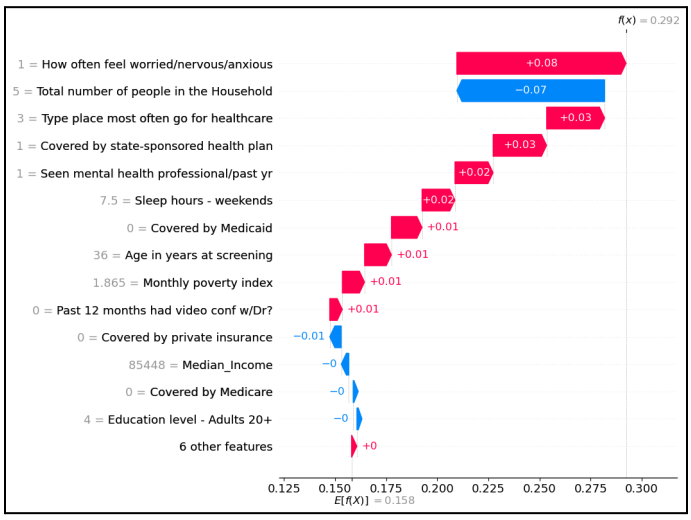


Figure 9. *SHAP values for a misclassified case (True = Moderate, Predicted = Severe) using the Random Forest model. Anxiety frequency and mental health access indicators increased the predicted severity.*

Model Evaluation and Results

All models followed the same tuning and evaluation pipeline. Random Forest achieved the highest macro-averaged ROC-AUC (0.588) and showed the broadest coverage across PHQ-9 severity classes. Logistic Regression yielded slightly higher overall accuracy but failed to correctly classify any Severe cases. SVM performed

comparably to Logistic Regression, with slightly lower precision in the upper severity categories.

As shown in Figure 6, the confusion matrices reflect class imbalance and difficulty distinguishing between adjacent categories. Random Forest outperformed the other models in identifying Moderately Severe and Severe cases. Figure 7 highlights that Random Forest relied on a combination of behavioral predictors (sleep and anxiety) and structural features (income and age), reinforcing the value of multi-domain inputs in depression classification.

Figure 8 provides a case-level explanation of one misclassified Moderate instance. Additional examples are included in Appendix B.

A reduced Random Forest model was trained using only socioeconomic features, including Median_Income, Monthly poverty index, education level, and health insurance type. This version followed the same cross-validation and evaluation pipeline as the full model.

Performance dropped substantially. Accuracy fell to 0.26, with a weighted F1 score of 0.25 and a macro-averaged ROC-AUC of 0.504. As shown in Figure 9, predictions clustered in the Mild and Moderate categories, and the model failed to correctly classify any Severe cases. Class sensitivity declined overall, confirming that behavioral and clinical variables such as sleep and anxiety were critical for improving prediction accuracy.

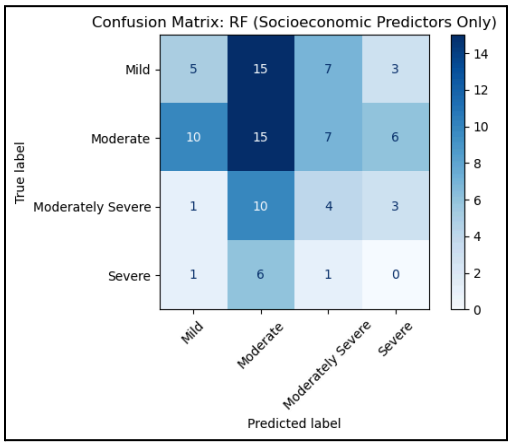


Figure 10. *Confusion matrix for the structural-only Random Forest model using socioeconomic predictors.*

Failure Analysis

The tuned Random Forest model consistently underpredicted Severe cases. One example involved an 18-year-old respondent with no regular healthcare access, a high poverty index, and public insurance. Despite these structural risk factors and a Severe PHQ-9 score, the model predicted Moderate.

SHAP values showed that anxiety frequency and sleep duration increased the predicted severity, but their impact was reduced by household size and insurance type. These opposing effects prevented the model from reaching the correct classification. This pattern highlights the difficulty of detecting high-risk individuals when symptom distributions overlap and class imbalance is present. Potential improvements could include ordinal classification, class weighting, or engineered interactions between social and behavioral features.

DBSCAN into supervised learning

This analysis examined the progression from unsupervised to supervised learning in modeling predictors of depression severity using PHQ-9 survey data, demographic attributes, and an engineered socioeconomic feature. The process began with the application of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a non-parametric unsupervised learning algorithm. After appropriate preprocessing—including the imputation of missing values and feature standardization—DBSCAN identified meaningful clusters within the dataset based on density. From this clustering, the fifteen most salient variables contributing to cluster formation were retained, serving as the foundation for the supervised learning phase.

To enrich the dataset with socioeconomic context, an estimated median income feature was engineered. This was calculated by associating each respondent's demographic profile (including gender, race, age, and education) with U.S. Census-based median income values for those categories. Each individual's estimate was computed by averaging the median incomes corresponding to their demographic segments. This feature was then appended to the set of predictors used in the supervised models.

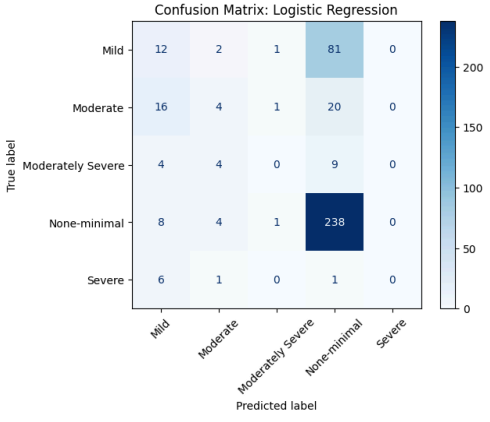


Figure 11. *Confusion matrix for the Logistic Regression model using socioeconomic predictors.*

Supervised learning techniques were applied using the PHQ-9 severity levels—categorized into None-minimal, Mild, Moderate, Moderately Severe, and Severe—as the target variable. A confusion matrix was generated to examine the relationship between input features and the PHQ-9 severity categories (Figure 10). This helped assess feature redundancy and detect initial patterns. Three supervised models were subsequently trained and evaluated: Logistic Regression, Random Forest, and Support Vector Machine (SVM). The dataset was split using stratified sampling to preserve the distribution of the outcome classes, and all features were standardized to support convergence and comparability across models.

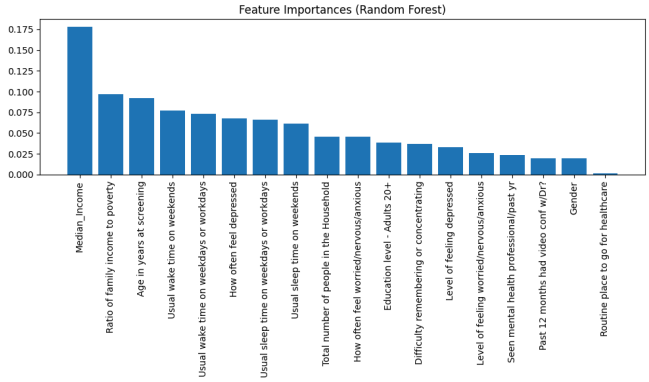
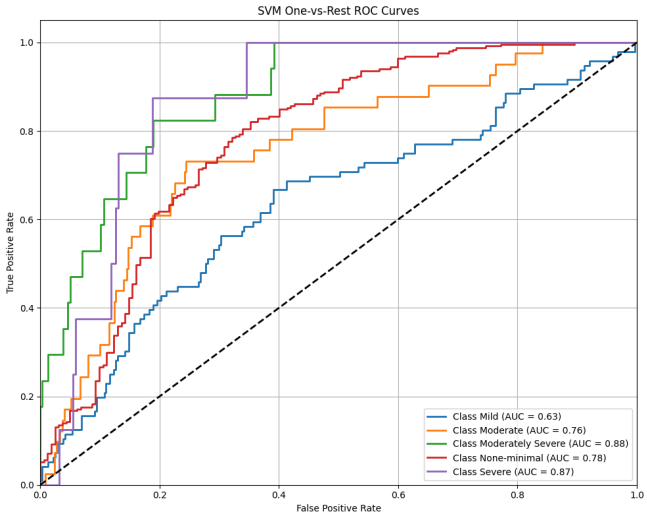


Figure 12. *Feature importance for the Random Forest model using socioeconomic predictors.*

Each model's performance was assessed using F1-score, overall accuracy, and macro-averaged area under the Receiver Operating Characteristic curve (ROC-AUC) using a one-vs-rest (OvR) framework. These metrics enabled a comprehensive understanding of how well each model distinguished between severity levels. To provide interpretability, variable influence was visualized through multiple techniques. Logistic Regression offered coefficient estimates along with 95% confidence intervals for each predictor, highlighting the direction and statistical significance of effects across severity levels. The Random Forest model's feature importance scores provided another view of variable influence (Figure 11), with variables such as estimated median income and several demographic features emerging as consistently influential.

Permutation importance was also computed for Random Forest to assess the stability and marginal impact of features on model accuracy. The resulting visual illustrated variability in each predictor's contribution through boxplots, adding another layer of interpretability. Since the Support Vector Machine model does not provide direct feature importance, its performance was illustrated using the ROC curve, offering a visual summary of the model's class-wise discrimination ability (Figure 12).

Figure 13. *One vs Rest ROC curves from SVM model using socioeconomic predictors.*



Together, these steps—from DBSCAN-based variable filtering to supervised model evaluation and interpretation—created a coherent and rigorous pipeline for understanding and predicting depression severity in population health data. The strategically placed visual outputs supported and enhanced each analytical phase: the heatmap (Figure 10) introduced the foundational structure of the feature-target relationship, the logistic regression confidence intervals offered inferential insight, the Random Forest feature importance (Figure 11) and permutation plots provided robustness checks, and the ROC curve for SVM (Figure 12) demonstrated end-to-end classification performance. This structured and interpretable transition from unsupervised clustering to supervised prediction reflects a comprehensive approach to modeling mental health outcomes using publicly available data.

Supervised Learned Based on PCA-Derived Predictors

This analysis was done using the top features that were produced from the PCA model that was made and analyzed earlier. Taking those features, another feature was added based on the Census data that combines demographic data to better inform the model. These features were used as the input for the 3 supervised learning models: Logistic Regression, Random Forest Classifier, and Support Vector Machines.

GridSearchCV and Stratified 5-Fold validation were used to select the hyper parameters for each model.

Model performance was evaluated using accuracy, weighted F1-score, and macro-averaged ROC-AUC score. This was done to see how the model performed using different metrics. On top of this a classification report was generated to see how well the model performed on each of the 5 classes.

After the three models were run and analyzed the best performing model was chosen, in this case it was Logistic Regression. After that model was chosen, a sensitivity analysis was done by removing all depression columns and only keeping the columns that related to socioeconomic factors. This was done to see how well the best model would be affected by keeping the socioeconomic variables as that is the focus of this project.

Model Evaluation

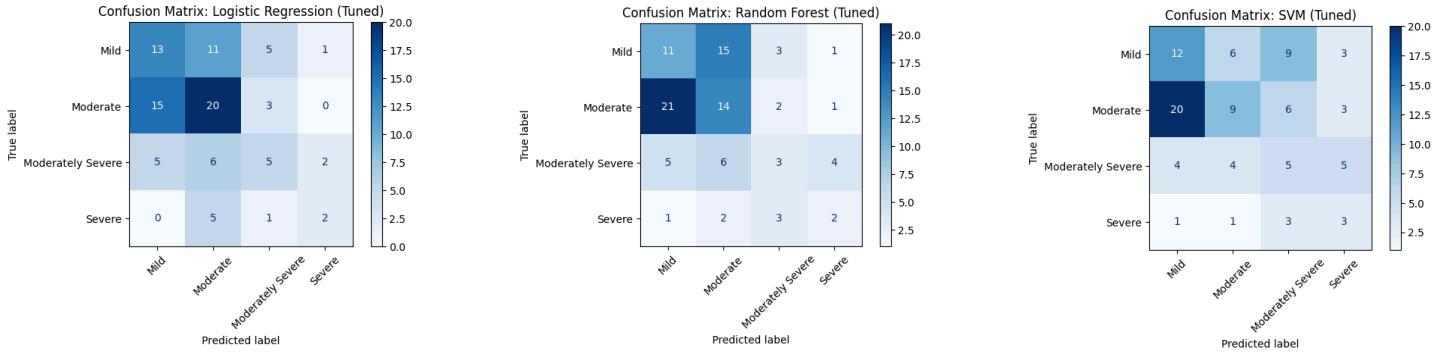


Figure 14. *The confusion matrices that show model performance over the three models. The graphs show performance by displaying predicted vs true models.*

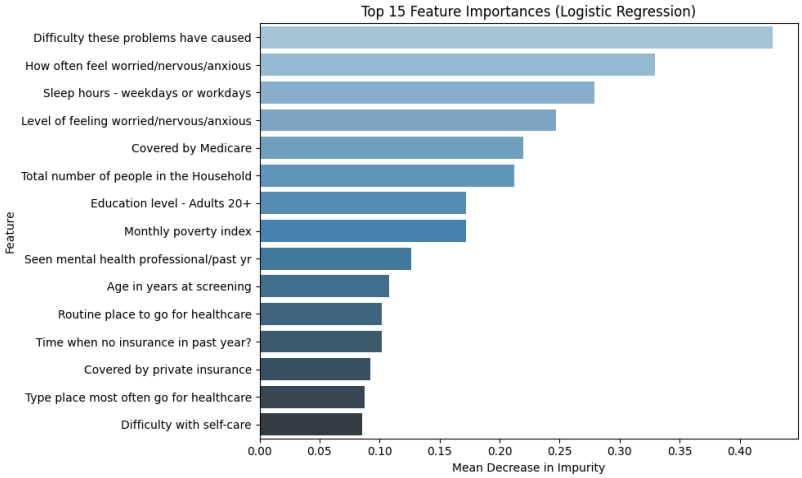


Figure 15. *The top 15 features were gathered after the Logistic Regression Model was trained. This model was chosen at it was the highest performing model in the PCA analysis*

Model Evaluation and Analysis

Overall, the model did not perform as well as expected. Most of the accuracy scores were around 0.4, F1-scores were around 0.36, and the ROC-AUC scores were around 0.6. These are low scores in each metric. As shown in Figure 14, the model has best success predicting the mild and moderate classes and struggled severely on the other ones. This could lend to the idea that the decision boundary of the model lies around between moderate and moderately severe classes. The best performing model was the Logistic Regression model that had an accuracy, f1-score and ROC-AUC score of 0.44, 0.40, and 0.663 respectively. Sensitivity Analysis was a technique that was used to see how well the model performed without the depression predictors

Sensitivity and Failure Analysis

The model that was selected for this analysis was Logistic Regression as it was the highest performing model. The features were selected by taking the top features from the PCA model and dropping all features that were not related to socioeconomic predictors. Overall, the model performance decreased when the depression variables were dropped as expected. The accuracy went from 0.44 to 0.36, F1-score went from 0.40 to 0.26 and the ROC-AUC did not change. This shows that the model had a significant decrease in performance during the sensitivity analysis.

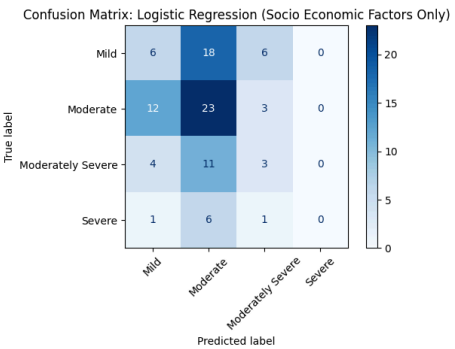


Figure 16. *This graph is the confusion matrix that was produced from the Logistic Regression model that only used socio economic factors.*

Performance Metrics for Logistic Regression, Random Forest, and SVM Models Informed by KMeans, PCA, and DBSCAN Analyses

	Accuracy	Weighted F1-Score	Macro-Averaged ROC-AUC
LogReg KMeans	0.32	0.30	0.568
LogReg PCA	0.44	0.40	0.663
LogReg DBSCAN	0.62	0.22	0.792
RF KMeans	0.31	0.31	0.588
RF PCA	0.32	0.30	0.628
RF DBSCAN	0.71	0.43	0.884
SVM KMeans	0.32	0.30	0.549
SMV PCA	0.40	0.36	0.661
SVM DBSCAN	0.63	0.22	0.785

Discussion

The unsupervised learning phase offered a valuable exploratory foundation for understanding the structure of the data before building predictive models. What stood out most during this phase was the overall consistency across different clustering techniques; the clusters that emerged were relatively stable regardless of the method used. A particularly notable insight was the prominence of sleep-related variables in differentiating groups. Variables such as average sleep duration and self-reported sleep trouble consistently appeared as distinguishing features across clusters, suggesting that sleep patterns may serve as early indicators or contributing factors to depressive symptoms. Interestingly, some of the expected leading indicators—such as frequency of feeling nervous or anxious—did not emerge as strongly within the clustering patterns, which highlighted the complexity and multifaceted nature of mental health beyond isolated symptoms.

In the supervised learning phase, the estimated income feature stood out as the most important predictor of depression. This was a significant finding, especially given that the feature was not directly reported by participants but derived through external Census data and demographic interpolation. Its importance reinforces the strong link between socioeconomic context and mental health outcomes and validated the effort to engineer richer, more informative predictors beyond what NHANES provides directly. The models also reaffirmed the importance of sleep-related variables and general health perception, both of which ranked highly in feature importance across multiple supervised models. These results not only aligned with prior research but also underscored the value of combining external data sources with standard survey variables.

Several limitations and challenges emerged throughout the process. One challenge involved balancing data completeness with model accuracy—some high-interest variables had high rates of missingness or inconsistent coding, requiring tradeoffs between sample size and variable richness. Additionally, although the estimated income feature proved powerful, it relied on generalizations from national-level Census data, which may not fully capture regional or local socioeconomic variation. If time and resources had allowed, future work could have dramatically expanded the geographic granularity of the analysis. A more detailed pipeline might link NHANES respondents to specific counties, states, or urban/rural designations using geographic metadata or external mapping tools. Doing so would allow for more contextually grounded estimates of income, cost of living, healthcare access, and social support structures—leading to more precise, actionable insights. Incorporating hyper-local demographics would also enable a deeper understanding of how place-based inequality intersects with mental health, which could ultimately inform more targeted policy and intervention strategies.

Ethical Considerations

One ethical issue that was encountered was how heavily we rely on the PHQ-9 survey. The survey in the case of this study is essentially used as a scale of how depressed someone is, when in reality there are many factors that are difficult to quantify that should be used.

There always exist ethical issues when working with health data even if it has been anonymized (as was done in this dataset.) This is because given enough personal information about a patient as given in the demographic data, there is always a chance of being able to be reidentified. This means that caution must be taken when compiling and working with highly sensitive data. There are also large ethical and legal consequences if this were to occur.

Another ethical consideration is how people could interpret the results of a model. For example, let's say the Random Forest model says that the most important feature to the target variable, 'PHQ-9 Total' is whether someone has health insurance or not. Someone could interpret this as meaning that they must get health insurance or else there is a high chance they get depression. This analysis by itself is incorrect as there are many factors to take into account such as model scores and how other variables are scored. It is important to include proper interpretations of the quantitative data as to not mislead people.

Statement of Work

Alexis Parker	Brandon Fox	Vikram
GitHub repo and script setup (Jupyter, Python), data cleaning and merging, KMeans clustering (UL), supervised modeling (Logistic Regression, Random Forest, SVM), evaluation (F1, ROC-AUC, SHAP), failure and sensitivity analysis, visualizations (matplotlib, SHAP), collaborative report writing	DBSCAN (UL), supervised modeling (Logistic Regression, Random Forest, SVM), evaluation (F1, ROC-AUC, SHAP), failure and sensitivity analysis, visualizations, collaborative report writing, census joining	PCA (UL), supervised modeling (Logistic Regression, Random Forest, SVM), evaluation (F1, ROC-AUC, SHAP), failure and sensitivity analysis, visualizations (matplotlib, seaborn), collaborative report writing

References

Califf, R. M., Wong, C., Doraiswamy, P. M., Hong, D. S., Miller, D. P., & Mega, J. L. (2021). *Importance of social determinants in screening for depression*. Journal of General Internal Medicine, 37, 2736–2743. https://doi.org/10.1007/s11606-021-06957-5

Kim, C. (2025). *Exploring factors influencing depression: Socioeconomic perspectives using machine learning analytics*. Electronics, 14(487), 1–15. https://doi.org/10.3390/electronics14030487

Vu, T., Dawadi, R., Yamamoto, M., Tay, J. T., Watanabe, N., Kuriya, Y., Oya, A., Tran, P. N. H., & Araki, M. (2025). *Prediction of depressive disorder using machine learning approaches: Findings from the NHANES.* BMC Medical Informatics and Decision Making, 25(83). https://doi.org/10.1186/s12911-025-02903-1

Appendix A. Extended Unsupervised Results: Feature Patterns and Comparisons

A.1 Top 15 Feature Patterns by Cluster (Imputed Version)

The imputed dataset (n = 467) retained all records by filling missing values using median or mode imputation, depending on the variable type. After assigning clusters with K-Means, average values were calculated for each feature within each cluster. The 15 features with the largest mean range across clusters were selected to highlight those most associated with subgroup separation.

Differences in age, poverty index, and PHQ-9 scores were among the most distinguishing. One cluster skewed older with lower PHQ-9 scores, while another had elevated symptom levels along with higher rates of telehealth use and alternative care settings. These patterns suggest that care access and behavioral differences may influence depression risk profiles in distinct ways.

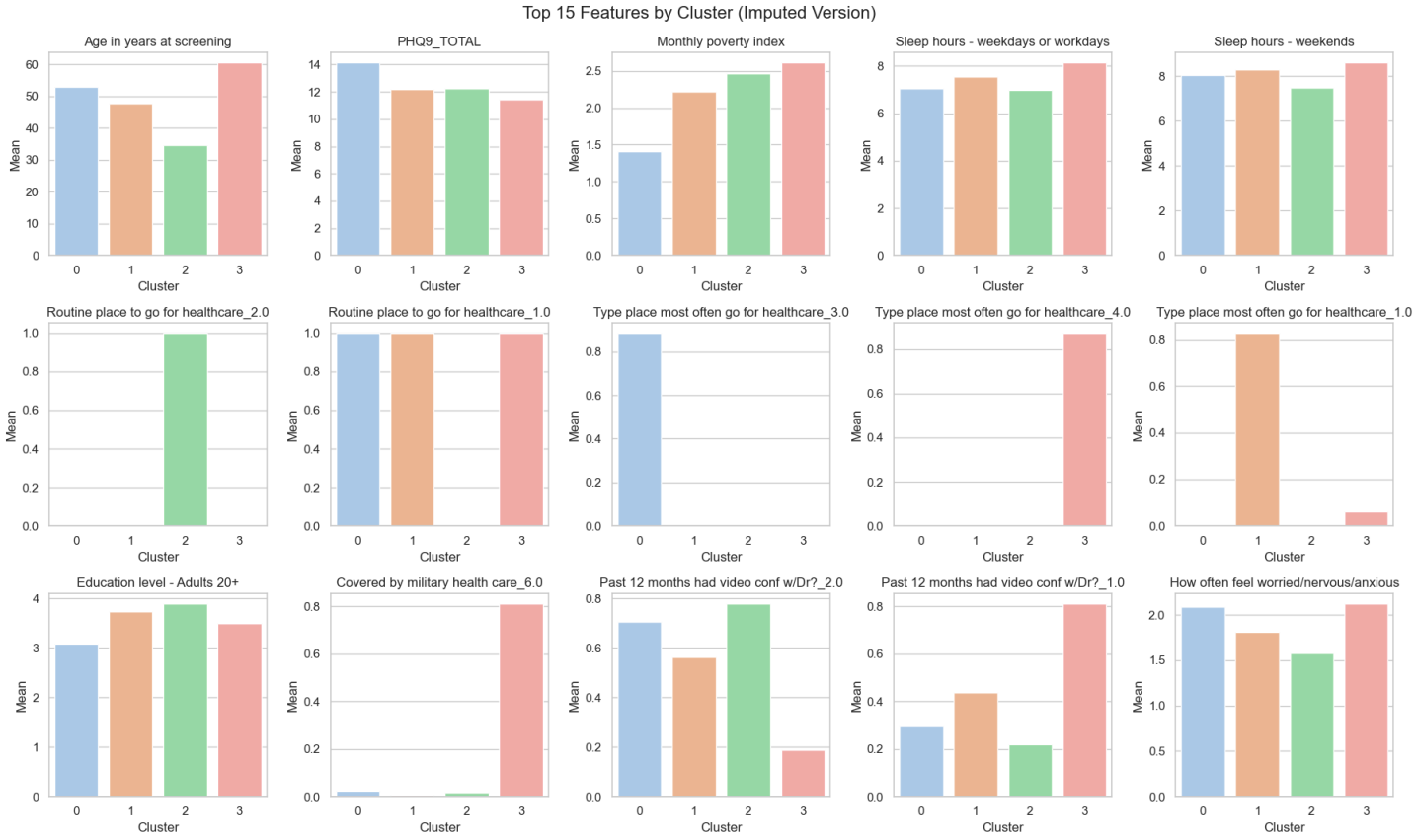


Figure A1. *Mean values of the top 15 features by cluster in the imputed dataset, ordered by within-feature range.*

A.2 Comparison of Top-Ranked Features Across Preprocessing Versions

Cluster results from the drop and imputed datasets were compared by ranking features according to the range of their mean values across clusters. A side-by-side comparison of the top 15 features from each version revealed both overlap and divergence. Several features appeared in both lists, including age, sleep duration, and indicators of healthcare access.

The imputed version included PHQ-9 scores among the top differentiators, which was expected because imputation preserved complete outcome data. In contrast, the drop version emphasized features such as visits to mental health professionals and insurance type, reflecting patterns that were more prominent in complete cases.

Rank	Drop Version Feature	Imputed Version Feature
1	Age in years at screening	Age in years at screening
2	Monthly poverty index	PHQ9_TOTAL
3	Sleep hours - weekdays or workdays	Monthly poverty index
4	Sleep hours - weekends	Sleep hours - weekdays or workdays
5	Seen mental health professional/past yr_2.0	Sleep hours - weekends
6	Seen mental health professional/past yr_1.0	Routine place to go for healthcare_2.0
7	Education level - Adults 20+	Routine place to go for healthcare_1.0
8	Past 12 months had video conf w/Dr?_1.0	Type place most often go for healthcare_3.0
9	Past 12 months had video conf w/Dr?_2.0	Type place most often go for healthcare_4.0
10	Time when no insurance in past year?_2.0	Type place most often go for healthcare_1.0
11	How often feel worried/nervous/anxious	Education level – Adults 20+
12	Routine place to go for healthcare_1.0	Covered by military health care_6.0
13	Routine place to go for healthcare_2.0	Past 12 months had video conf w/Dr?_2.0
14	Type place most often go for healthcare_1.0	Past 12 months had video conf w/Dr?_1.0
15	Covered by private insurance_1.0	How often feel worried/nervous/anxious

Figure A2. *Comparison of features ranked by importance in drop and imputed cluster results.*

A.3 Heatmap of Feature Overlap

A binary matrix was used to visualize the overlap between the top 15 features from each clustering result. Rows represent features from the drop version, and columns represent those from the imputed version. Green squares indicate features that appeared in both lists.

The presence of shared variables such as age, weekday and weekend sleep hours, and healthcare usage indicators reinforces the consistency of key patterns across preprocessing approaches. While some variation exists based on how missing data was handled, the core structure of important features remained stable.

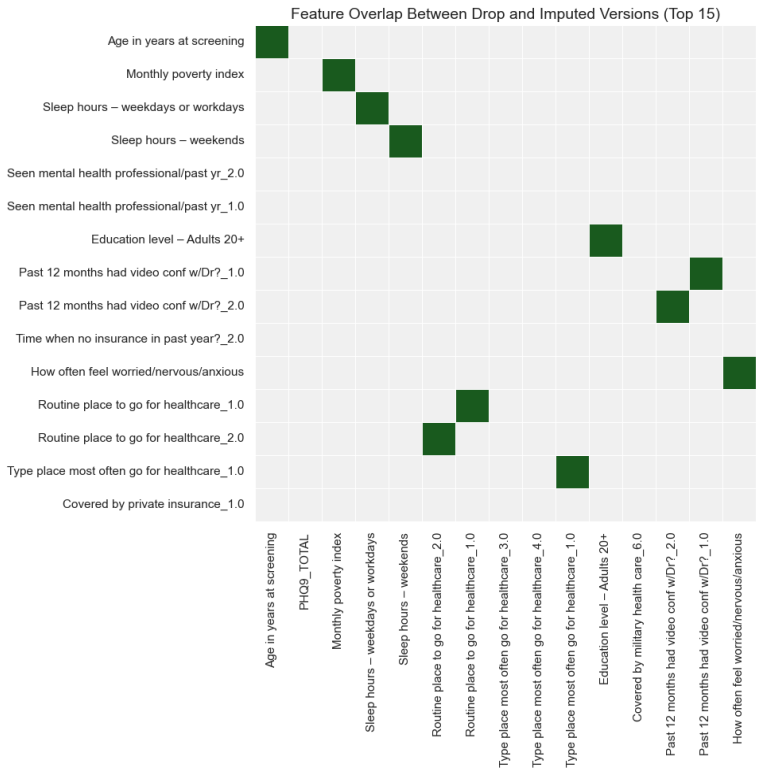


Figure A3. *Heatmap showing feature overlap between the top 15 features from drop and imputed cluster results.*

Appendix B. Extended Supervised Results: SHAP Explanations and Sensitivity

B.1 SHAP Explanations for Selected Cases

This section includes additional SHAP waterfall plots from the tuned Random Forest model using KMeans-derived predictors. Each plot illustrates how individual features increased or decreased the model's predicted severity for selected respondents.

Figure B1. SHAP – Misclassified Moderate Case (Predicted Mild)

A 71-year-old respondent was labeled Moderate but predicted as Mild by the model. Anxiety symptoms and a higher poverty index contributed to an upward adjustment in risk. However, older age, Medicare coverage, household size, and minimal sleep-related variation reduced the overall score. The combined effect of weakly contributing features and downward-pulling factors resulted in underprediction despite signs of elevated risk.

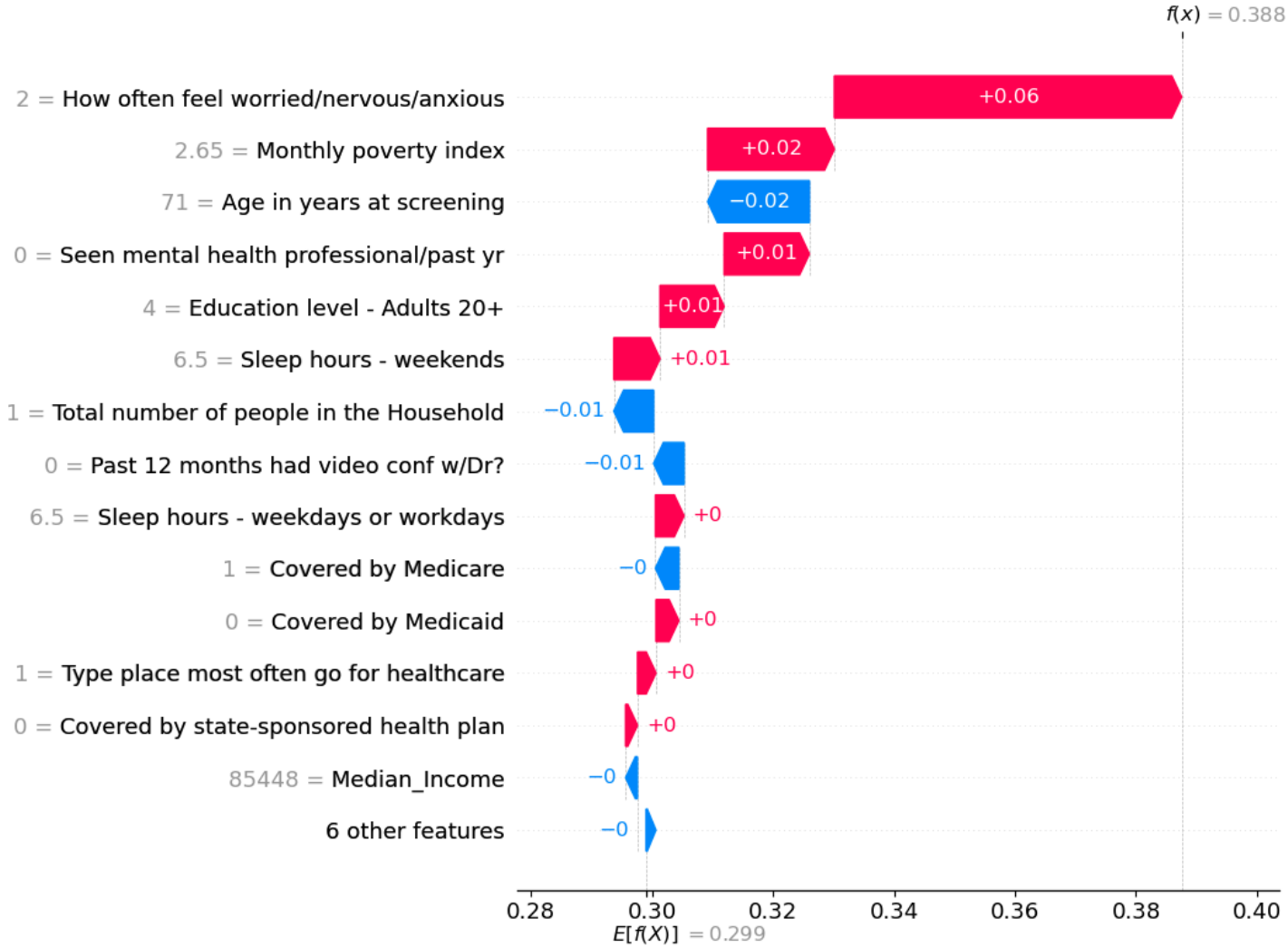
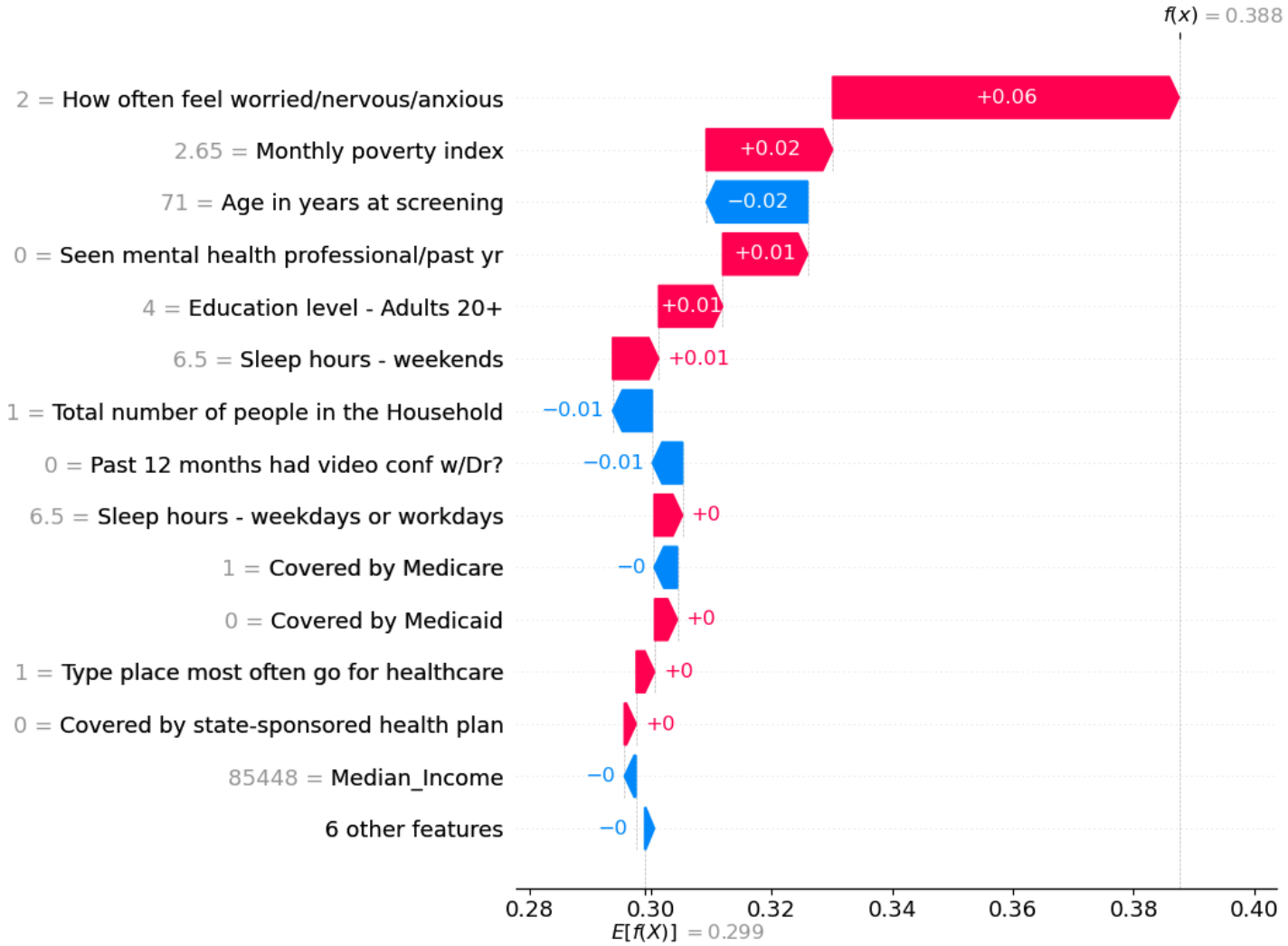


Figure B2. SHAP – Misclassified Moderately Severe Case (Predicted Mild)

This borderline case involved multiple muted contributions. Education level, poverty index, and income increased the predicted risk. However, anxiety frequency, weekday sleep, and private insurance coverage pulled the score downward. The overall feature mix lacked a dominant signal, and key indicators like anxiety and access to care did not shift the prediction far enough to reach the correct severity. These offsetting effects led to underclassification.


f(x) = 0.388
2 = How often feel worried/nervous/anxious
2.65 = Monthly poverty index
71 = Age in years at screening
0 = Seen mental health professional/past yr
4 = Education level - Adults 20+
6.5 = Sleep hours - weekends
1 = Total number of people in the Household
0 = Past 12 months had video conf w/Dr?
6.5 = Sleep hours - weekdays or workdays
1 = Covered by Medicare
0 = Covered by Medicaid
1 = Type place most often go for healthcare
0 = Covered by state-sponsored health plan
85448 = Median_Income
6 other features
+0.06
+0.02
−0.02
+0.01
+0.01
+0.01
−0.01
−0.01
+0
−0
+0
+0
+0
−0
−0
0.28
0.30
0.32
0.34
0.36
0.38
0.40
E[f(X)] = 0.299