# The Evaluation of How Healthcare Access Impacts Self-Reported Mental Health Outcomes – Initial Models Report

GitHub Link: https://github.com/stevensps/DSE6311-DataScience-Capstone

Benjamin Kelley*

Christopher Rodgers*

Steven Simonsen*

*All team members contributed equally to all aspects of this project.

## Abstract

This project uses the 2023 Behavioral Risk Factor Surveillance System (BRFSS) dataset to evaluate how healthcare access predicts self-reported mental health outcomes. The primary goal was to determine whether individuals with limited healthcare access report more frequent poor mental health days. To do this, we transformed the original response variable ment14d into a binary outcome: Low (0–13 days) vs. High (14+ days), aligned with CDC standards. We applied binomial logistic regression, Random Forest, and XGBoost models to predict this binary outcome, with Random Forest performing best. We found that factors like having a primary care doctor, insurance coverage, struggles with healthcare costs, and personal experiences with mental health challenges all play a significant role in predicting mental health outcomes. These insights can help us understand who might be at risk and support efforts to shape more effective public health strategies.

## Introduction

As mental health continues to pose a pressing need as a public health concern in the United States, significant implications for individual well-being, healthcare systems, and social productivity are of prominent concern. Understanding the factors that contribute to the level of an individual's mental health is essential for shaping effective interventions and informing policy. This project leverages data from the 2023 BRFSS dataset, one of the largest, continuously conducted health-related surveys in the world, collecting self-reported information on health behaviors, conditions, and access to care from over 400,000 U.S. adults annually (CDC 2023).

The BRFSS dataset's breadth and granularity offer an opportunity to explore complex health phenomena such as self-reported mental health status through social, behavior and access related variables. We focused on the variable ment14d, which measures the number of poor mental health days out of the past 30, and transformed it into a two-level categorical outcome—Low and High—to support a nuanced binary classification approach. Our decision to dichotomize our response variable (ment14d) into a newly created variable, titled ment14d_cat, is supported by public health practices in public health research. Per the CDC, frequent mental health distress is defined as experiencing 14 or more days within the past month, and this measure is utilized in population-based studies to assess mental health outcomes (Blackwelder, Amanda, et al., 2021).

By modeling self-reported mental health status using features such as assigned PCP, insurance coverage, personal healthcare access, and other demographic variables, this analysis aims to uncover patterns that could support early identification of at-risk individuals and highlight systemic gaps in healthcare accessibility. Our goal is to allow our findings to inform public health campaigns and healthcare resource allocation by clarifying the predictive relationships between healthcare access and mental health. There is also an opportunity to use this model as a predictor to incorporate into a health screen at doctors appointments. This would allow healthcare providers to ask a quick question about mental health which could initiate a longer conversation if necessary. Ultimately, this work contributes to the growing field of mental health data science, using machine learning and statistical modeling to turn survey data into actionable insights.

## Background

In more recent years, mental health has come to the forefront of many conversations surrounding the overall well-being of those residing in the United States. According to the National Institute of Mental Health (2024), an estimated one in five U.S. adults are living with a mental illness. Furthermore, access to healthcare due to cost has also been a popular topic of discussion. In 2023, 63% of uninsured adults ages 18-64 said that they were uninsured because the cost of coverage was too high (Cervantes, Tolbert, 2024). Therefore, the research conducted in this project aims to answer the following question: Can we predict poor mental health outcomes, measured by self-report, among Americans by access to healthcare, after controlling for mental illness, demographics and other exposures?

By using the Behavioral Risk Factor Surveillance System (herein referred to as the BRFSS dataset), our research will examine and determine best practices for conducting machine learning (ML) and predictive modeling using the BRFSS dataset or other similarly large social surveys. The goal of this data science analysis is to use our findings from the unweighted dataset to contribute to a larger project by setting the stage for future work in this area of research.

## Hypothesis and Prediction

Our hypothesis is that healthcare access has a significant bearing on self-reported mental health outcomes because individuals without sufficient access are likely to suffer from unresolved mental health issues, along with higher stress levels and lower utilization of essential health services. As such, we predict that machine learning models trained on BRFSS data will identify access-to-care variables as strong predictors of the number of self-reported poor mental health days, and individuals without access to healthcare will report a higher number of poor mental

health days. This question, and report, is directed to Social Science & Health Researchers across the U.S.

We predict that machine learning models trained on BRFSS data will identify access-to-care variables as strong predictors of the number of self-reported poor mental health days, and individuals without access to healthcare will report a higher number of poor mental health days. Additionally, we predict that mental health-related demographic factors such as diagnosed mental health illness, will contribute to the overall predictions made by the machine learning models.

## Data

The dataset used in this analysis is the 2023 Behavioral Risk Factor Surveillance System (BRFSS), obtained directly from the Centers for Disease Control and Prevention (CDC). BRFSS is one of the world's largest continuously conducted health surveys, gathering data via telephone from over 400,000 U.S. adults annually.

**Justification for Dataset Choice:**

The BRFSS dataset is highly appropriate for this project because it contains self-reported data on physical and mental health, healthcare access, chronic conditions, and health-related behaviors. It includes detailed variables that allow for analyzing both mental health outcomes and healthcare accessibility. The BRFSS dataset also supports large-scale, generalizable insights across U.S. regions due to its national scope and sample size. It is widely used in public health research and policy planning, making it relevant for stakeholder application.

**Dataset Overview:**

- **Dimensions:** Approximately 425,000 rows (individual respondents) and 22 selected variables after cleaning and feature selection.
- **Response Variable:**
    - _MENT14D: Original variable indicating the number of poor mental health days in the past 30.
    - ment14d_cat: A newly derived binary variable categorizing respondents as:
        - **Low**: 0–13 poor mental health days
        - **High**: 14+ poor mental health days
        - This transformation follows CDC definitions for "frequent mental distress" and supports binary classification modeling.

- **Predictor Variables Include:**
    - Access to healthcare (e.g., has insurance, has a primary care provider, cost-related care barriers)
    - Existing mental health conditions (e.g., diagnosed depression)
    - Cognitive limitations
    - Demographics (age, income, household composition, region)

**Known Issues and Limitations:**

- **Class Imbalance:** Most respondents fall into the "Low" category, requiring threshold tuning and up/down-sampling strategies to avoid bias toward the dominant class.

- **Self-Reported Data:** Responses may include recall bias or social desirability bias, which can affect the accuracy of both predictor and outcome variables.
- **Missing Values:** Several variables contained nonresponse options such as "Not Sure," which had to be cleaned or imputed.
- **Age Imbalance:** One of our key predictors, age, was skewed to represent mainly an older population. This is discussed in more detail in the Data Exploration section below.

**Survey Design Not Used:** This analysis uses unweighted BRFSS data, which doesn't account for the complex survey design. As a result, findings are not nationally representative, but still meaningful for modeling purposes.

We took several cleaning steps with the 2023 BRFSS dataset to prepare it for modeling and ensure everything was consistent and reliable. Below is a step-by-step account of our data cleaning process followed by a brief description of each step.

**1. Handling Missing Values:**

- Responses such as “Not sure” were recoded as NA.
- Rows with missing values in the target variable (ment14d_cat) were dropped entirely, since imputation would compromise the validity of the outcome.
- For all other variables, missing values were imputed using MICE via the futuremice implementation. This allowed for realistic estimation of missing values while preserving relationships among variables. Furthermore, the use of futuremice allowed for parallel processing across multiple CPU cores to reduce computation time.

**2. Transformations and Derived Variables:**

**Target variable transformation:**

- Original _MENT14D variable (number of poor mental health days in the past 30) was converted into a binary variable called ment14d_cat:
    - **Low** = 0–13 days
    - **High** = 14+ days
    - This aligns with CDC definitions for frequent mental distress and supports binary classification.

**Collapsing geographic categories:**

- The BRFSS state-level location variable was transformed into four U.S. Census regions:
    - Northeast, South, Midwest, West
    - This made modeling more interpretable and reduced category sparsity.

**Dummy encoding:**

- All categorical predictor variables were one-hot encoded, excluding the target variable. This allowed for compatibility with machine learning algorithms like Random Forest and XGBoost.

**3. Assumptions Made:**

- Treating “Not sure” and similar responses as missing values was based on the assumption that they do not provide reliable data for inference.
- By collapsing states into regions, we assumed regional trends are more relevant than state-level idiosyncrasies for mental health analysis.
- Missing at random (MAR) was assumed for imputation purposes, meaning that the probability of missingness depends on observed data, not unobserved.

**Replicability Notes:**

- These cleaning steps were implemented in R using packages like dplyr, tidyr, and futuremice.
- Anyone replicating this process should ensure to:
    - Use BRFSS 2023 raw .xpt data
    - Recode missing-like strings as NA
    - Perform imputation after one-hot encoding (as done here)
    - Drop NA rows only for the outcome variable

## Data Exploration

To begin our exploration, we wanted to explore how our pre-transformed response variable, ment14d, was distributed across the BRFSS. As seen below in Figure 1, the majority of the respondents had 0 poor self-reported mental health days in the last 30 days. We knew we had to address this imbalance, especially considering our decision to collapse this variable into ment14d_cat, which split the dataset into Low (0-13 poor self-reported mental health days), and High (14+ poor self-reported mental health days).

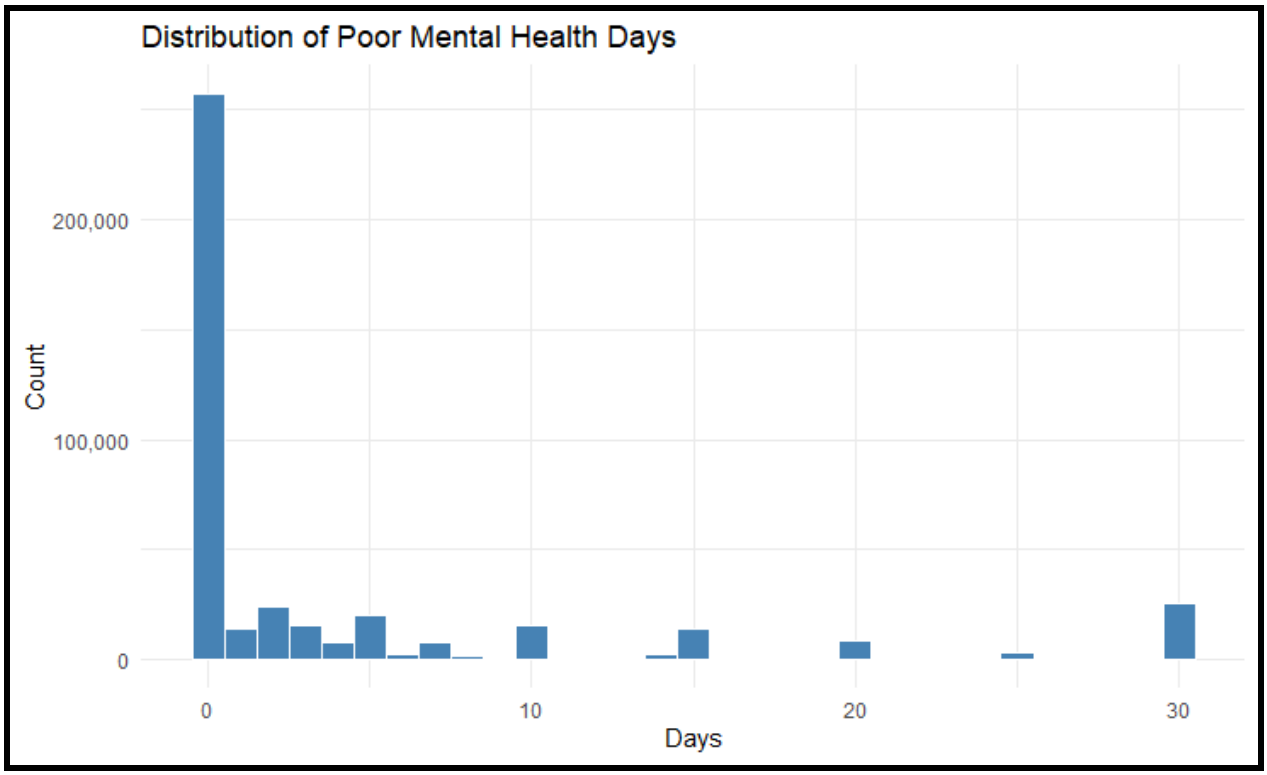


***Figure 1:*** *A breakdown of our pre-transformed response variable by number of poor self-reported mental health days in the last 30 days.*

To support our question and hypothesis, and highlight the issues in the Data Acquisition process, we wanted to illustrate Figure 2 below. This graph, while simple, highlights a few key facts about the dataset. First, as previously mentioned, these data are very much imbalanced. Initially, this not only led to poor computational performance, but also provided a recall value that was lower than desired. To improve performance both from a computational and modeling perspective, we downsampled the majority class (Low). The red bars graph below illustrate the total number of Low and High observations contained in the entire dataset. The blue bars, by contrast, represent the training dataset after processing the data via downsampling. As can be seen, the two classes are now equal to one another.

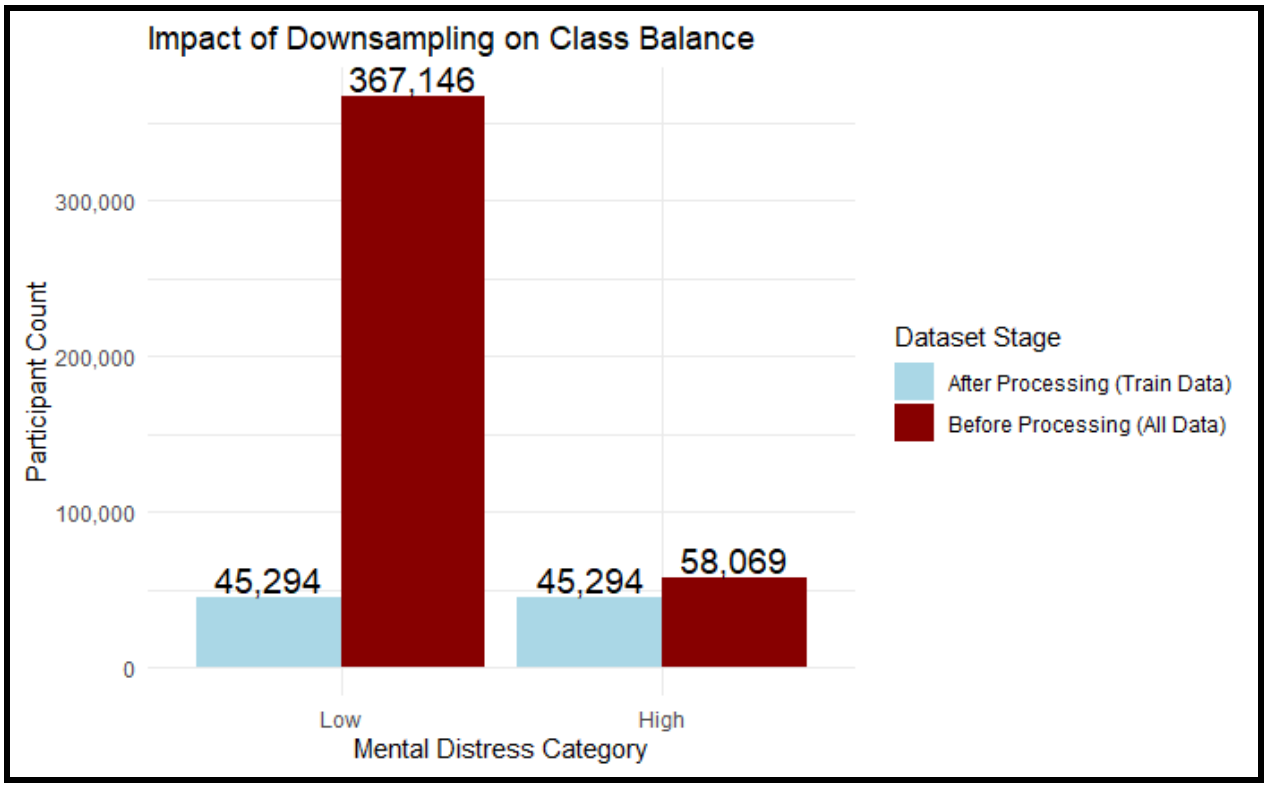


***Figure 2:*** *A bar graph showing the distribution of Low (0-13 bad mental health days in the last 30 days) and High (14+ bad mental health days in the last 30 days) in the original dataset, and the training data after downsampling.*

To further explore the dataset, we wanted to understand how various age groups were measured. In Figure 3 below, there are a few takeaways. First, the proportion, or percentage, of "Low" is much greater than "High" across all age groups. However, the younger the age group, the higher the proportion of High observations. Second, these data have significantly more observations for survey participants 60+ years old (202,718 observations in total). This explains an additional limitation of our model, which is that the resulting models may explain and predict behaviors best for a population 60 years old and above. In our future work, this is something we would like to explore in greater detail to understand the impact of this finding.

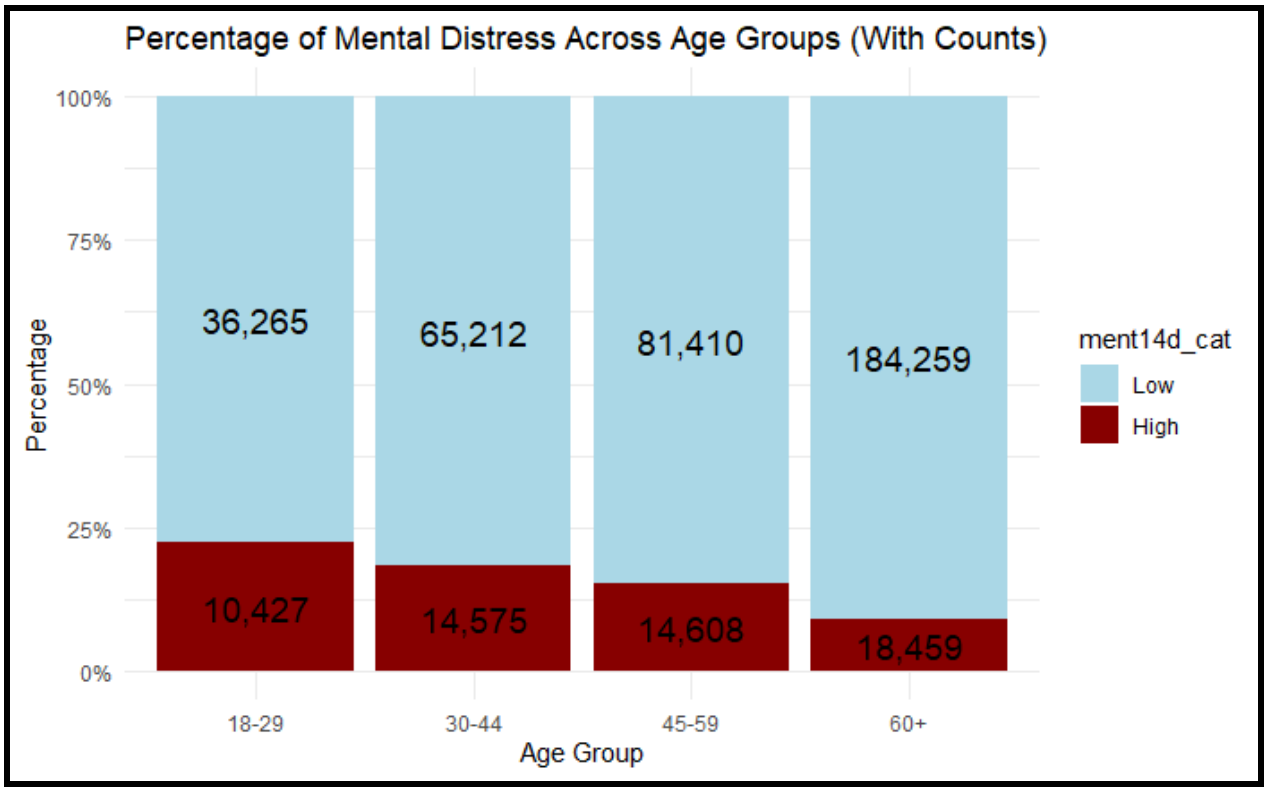


***Figure 3:*** *A stacked bar chart showing the count of Low and High for the various age groups contained in the dataset prior to encoding and imputation. The percentage of Low and High for each age group is shown on the y-axis.*

Figure 4, displayed below, shows the missingness of the dataset we used for modeling prior to imputation or encoding. While some variables are not missing at all, others are missing a sizable number of observations. Of note is our response variable, which displayed 8,108 missing observations. As previously noted, this was a small percentage of the overall dataset (1.9%), so we chose to drop these values to maintain the integrity of our response variable. Another interesting finding is the large number of missing variables in acedeprs (living with somebody depressed or suicidal), numadult (number of adults living in each household), and pimins1 (current source of primary health insurance). Relative to the total number of observations within each variable, 35% or more of the rows in each were missing. Futuremice was used to impute

values, however it may be worth conducting further analysis to determine how the large amount of missing data in these variables impacted the results of the research.

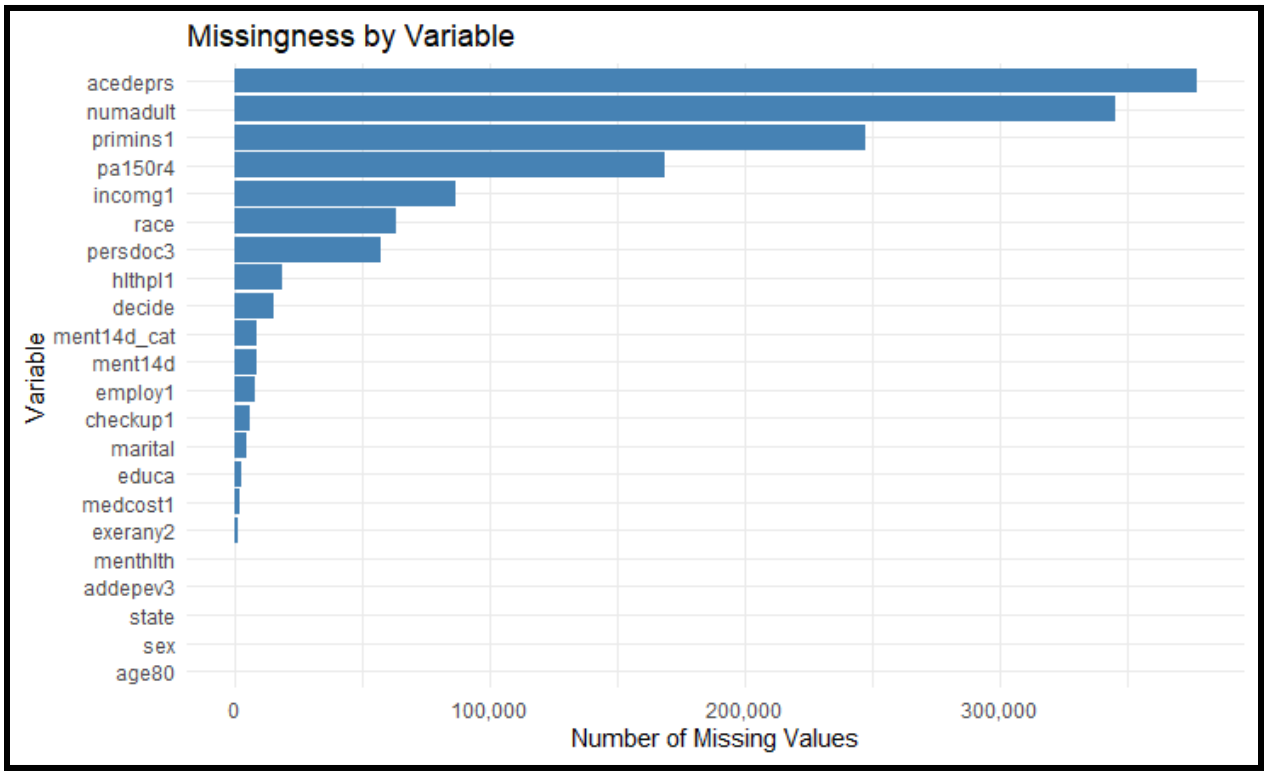


***Figure 4:*** *A graphical display of the number of missing values for each variable included in our research.*

Another area of interest in our data exploration includes the distribution of participants by income group, and how this played into whether or not they could see a doctor. As can be seen in Figure 5, the majority of each income group does not have a cost barrier for seeing a doctor in the last 12 months. However, there does seem to be a relationship between the amount of money an individual earns, and whether or not there is a cost barrier. For example, in the lowest income group (less than $10k), about 20% of the survey respondents reported avoiding the doctor due to cost. This ties back to our original hypothesis and prediction, illustrating that access to care is determined by demographic factors, such as income.

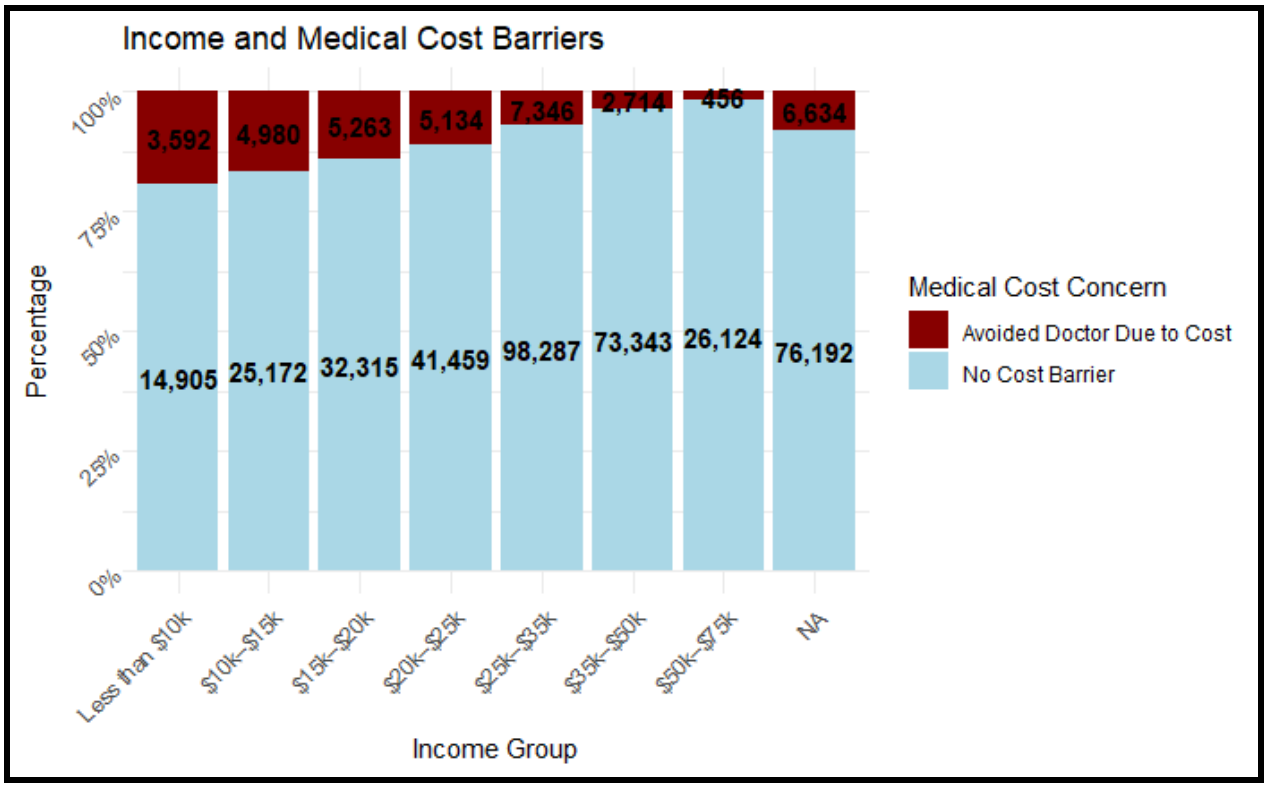


**Figure 5:** *Plot showing whether participants avoided the doctor due to cost by income group.*

## Models

### Preprocessing and Dimensionality Reduction

To prepare our data for modeling, we went through several preprocessing steps:

### Train/Test Split

We created stratified train/test splits to maintain the proportion of our target variable (ment14d_cat). This ensured the "Low" and "High" groups were fairly represented during training and evaluation.

### Missing Data Imputation

We first recoded responses like "Not Sure" as NA. Then, using MICE via the futuremice package, we imputed missing values across all applicable variables. One final manual fix was applied to the primins1_88 variable using mean substitution, which filled remaining gaps post-imputation.

**Categorical Encoding**

All categorical variables were one-hot encoded, except for the outcome variable. This format allowed our models—especially Random Forest and XGBoost—to work without assumptions about variable distributions.

**Transformations**

Since tree-based models don't require scaling, we did not apply any normalization or standardization beyond what was needed for PCA.

**Dimensionality Reduction**

We performed Principal Component Analysis (PCA) on the numeric training data (after cleaning and imputation) to assess whether dimensionality reduction could help uncover patterns in the data:

- Based on the scree plot and PCA summary, PC1 and PC2 explained about 9% of the total variance (PC1: ~5%, PC2: ~4%).
- A plot of PC1 vs. PC2, colored by mental health group, showed some overlap but no clear separation between the "Low" and "High" categories.

- To explore the structure further, we applied k-means clustering (k = 3) to the first two principal components. While the clusters didn't perfectly align with the actual categories, they did show partial grouping: for instance, Cluster 2 included over 26,000 of the High group, suggesting some weak structure.

Despite these insights, we ultimately did not use PCA components for modeling, as they didn't improve class separation or predictive clarity over the original features.

**Algorithm(s) Selection**

We tested a few supervised learning models to predict poor mental health outcomes (ment14d_cat). The primary algorithms we used were **Binomial Logistic Regression**, **Random Forest**, and **XGBoost**. Our goal was to find a model that not only performed well but could also handle the complexity of our dataset and offer meaningful insights.

**Supervised Algorithms Used**

- **Binomial Logistic Regression** – Used as a baseline model for comparison.
- **Random Forest** – A tree-based ensemble method that became our best-performing model.
- **XGBoost** – A gradient boosting framework that we explored to compare against Random Forest.

**Justification for Each Choice**

- **Logistic Regression** was chosen as a baseline due to its simplicity and interpretability. It makes assumptions about linear relationships between features and the log-odds of the outcome.
- **Random Forest** was selected because it handles non-linear relationships well, is robust to multicollinearity, and doesn't require feature scaling. It also provides variable importance scores, which helped us interpret the most influential predictors.
- **XGBoost** was included for its ability to optimize model performance through boosting, though it requires more tuning and regularization than Random Forest.

**Method Assumptions**

- **Logistic Regression** assumes independence of predictors, absence of multicollinearity, and a linear relationship between predictors and the log-odds of the response. While we did not explicitly test these assumptions (e.g., with VIF or diagnostic plots), the nature of our dataset, which included imputed values, one-hot encoded variables, and complex interactions, made it likely that some assumptions may not have held. Because of this, we treated logistic regression as a baseline model to benchmark performance rather than a candidate for final deployment.
- **Random Forest and XGBoost** make fewer strict statistical assumptions, which makes them more appropriate for our dataset, especially with the mix of categorical and imputed features.

**Overfitting Control**

To control for overfitting:

- We used stratified 5-fold cross-validation to maintain class proportions across folds.
- We used downsampling on the majority class ("Low") to improve class balance.
- We limited model complexity using hyperparameters like min_n and mtry.
- We tracked multiple evaluation metrics (PR-AUC, recall, precision, kappa, accuracy). We selected the best model based on recall, since our goal was to correctly identify individuals in the "High" mental health risk category.

We also plotted tuning results using autoplot (rf_results) to inspect model performance across combinations visually.

Comparative Model Metrics

| model | accuracy | kap | recall |
|---|---|---|---|
| Logistic Regression (No Tuning) | 0.518 | 0.152 | **0.911** |
| Random Forest | 0.476 | 0.128 | **0.919** |
| XGBoost | 0.787 | 0.149 | **0.293** |

***Table 1:*** *A tabular view of model performance.*

The table above summarizes the performance of three key models tested during our project: Logistic Regression, Random Forest, and XGBoost. We evaluated each model based on three main metrics: accuracy, Cohen's kappa (kap), and recall, to get a picture of their strengths and limitations given the nature of our data and target variable. We chose to put the most weight on

recall performance since identifying individuals who are at risk of poor mental health is our primary objective.

Among these models, XGBoost demonstrated the highest accuracy at 78.7%, outperforming both Logistic Regression (51.8%) and Random Forest (47.6%). This suggests that XGBoost was the most effective model overall at correctly predicting both positive and negative outcomes. However, when examining recall, a critical metric for our problem as mentioned above, Random Forest and Logistic Regression both outperformed XGBoost significantly. Specifically, Random Forest achieved a recall of 91.9%, and Logistic Regression achieved 91.1%, compared to the relatively low recall of 29.3% obtained with XGBoost.

This pattern highlights a key trade-off between the models: while XGBoost predicts overall outcomes more accurately, it fails to capture a large proportion of true positive cases, which could be a serious limitation depending on our stakeholder's priorities. This was an issue our team struggled with and continued to improve upon throughout this project. Conversely, Random Forest and Logistic Regression models, despite lower overall accuracy, capture a much higher proportion of the true positives, making them potentially better choices when minimizing false negatives is essential. Cohen's kappa scores, which adjust for chance agreement, were relatively low across all models, but again highest for XGBoost (0.149), suggesting moderate but not substantial agreement beyond chance.

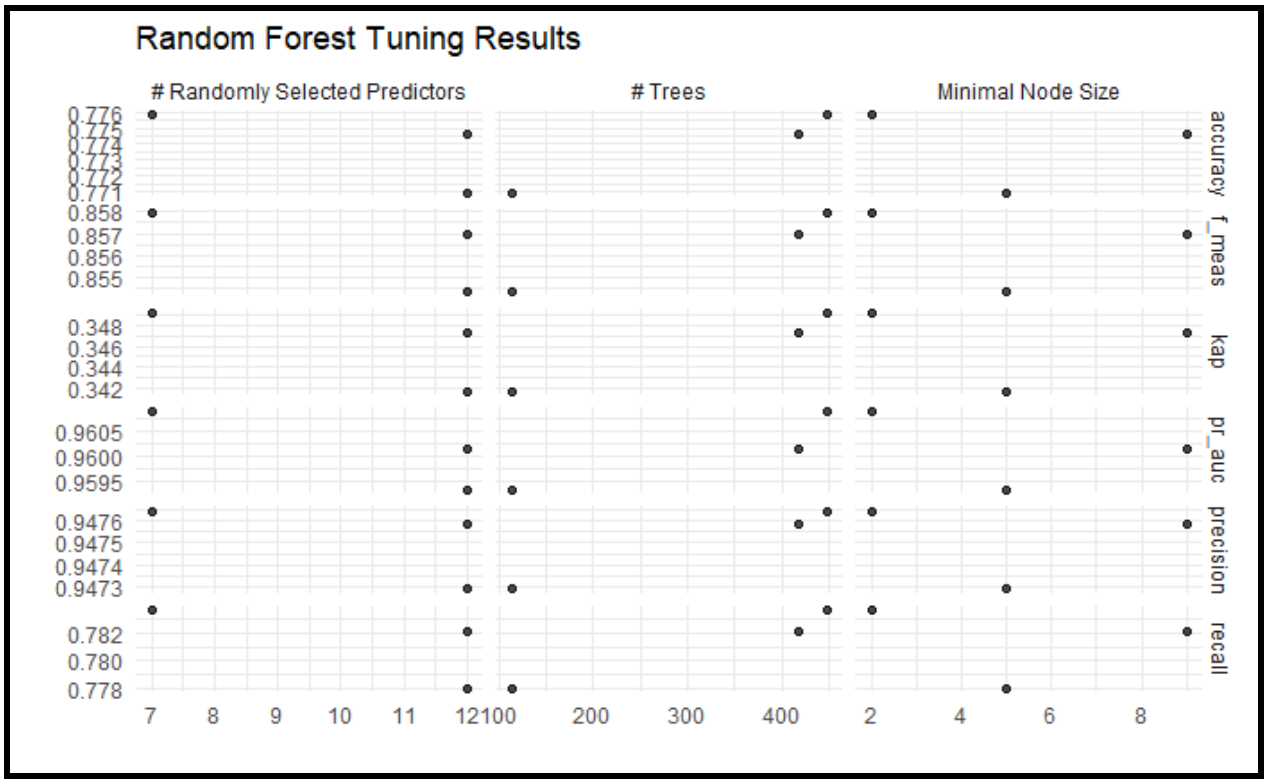


***Table 2:*** *The above table shows how the different characteristics of the random forest model affect performance metrics.*

Table 2 visualizes the performance outcomes across different configurations of Random Forest hyperparameters. Each dot corresponds to a model generated with a unique combination of:

- **Randomly Selected Predictors**: The number of features randomly selected at each tree split, ranging from 7 to 12 predictors.
- **Trees**: The total number of decision trees built in the forest, with tested values of 100, 200, 300, and 400 trees.
- **Minimal Node Size**: The minimum number of samples required to form a terminal node, with values from 2 to 8.

Model performance was then evaluated across 6 key metrics, including:

- **Accuracy:** The measure of proportion of all predictors that were correct, for both positive and negative classes.
- **F1 Score:** This measures the balance between precision and recall - a helpful metric when looking for a well-rounded model.
- **Kappa Statistic (KAP):** The measure of performance when compared to random guessing.
- **Area Under the Precision-Recall Curve (PR-AUC):** This summarizes the tradeoff between precision and recall. For values between 0 and 1, with the best result being as close to 1 as possible.
- **Precision:** This measures the proportion of positive predictions that were actually correct.
- **Recall:** Converse to precision, recall measures the proportion of actual positives that were predicted correctly.

As stated above, we assigned a greater importance to recall, although we took all metrics into account since ideally, we wanted a well-rounded model. The dot plot shows that varying the number of trees and minimal node size had noticeable effects on model metrics. Models with around 400 trees and minimal node sizes between 2 and 4 tended to yield higher accuracy and precision. Increasing the number of randomly selected predictors did not consistently improve model performance, suggesting that careful control of feature randomness was necessary. Recall values remained relatively high across multiple configurations, reflecting the model's strength in correctly identifying positive cases.

In summary, model selection depends heavily on whether the priority is maximizing overall correct predictions (favoring XGBoost) or ensuring high sensitivity to positive cases (favoring

Random Forest). Further tuning or hybrid approaches could be considered to attempt improving both recall and overall accuracy simultaneously. Overall, this tuning process demonstrated that while Random Forest models are robust, achieving the best performance required balancing tree depth, the number of trees, and the degree of randomness in feature selection.

## Discussion & Next Steps

### Summary of Key Takeaways

This project aimed to determine whether poor mental health outcomes, measured by self-reported mental health days, can be predicted by healthcare access among U.S. adults, after accounting for mental illness, demographics, and other exposures. Using data from the 2023 Behavioral Risk Factor Surveillance System (BRFSS), we created a binary target variable (ment14d_cat) that aligned with CDC standards, classifying individuals as either "Low" (0–13 days) or "High" (14+ days) based on the number of poor mental health days reported in the past 30.

We performed data cleaning and imputation to answer our research question, recoded "Not Sure" responses as missing and applied multiple imputations via MICE. After dummy encoding categorical variables and splitting the dataset into stratified training and test sets, we applied three supervised learning models: Binomial Logistic Regression, Random Forest, and XGBoost. We tuned each model using stratified 5-fold cross-validation and evaluated performance across several metrics. Model selection was based on recall, prioritizing correctly identifying individuals at higher mental health risk.

The Random Forest model performed best, achieving 93% recall for the High category using a custom threshold of 0.25. Top predictors included depressive disorder, cognitive difficulties, living with someone experiencing mental health issues, and age. Notably, several healthcare access variables—such as lack of insurance, difficulty affording care, and not having a primary care provider—emerged as strong predictors of poor mental health outcomes.

Our hypothesis stated that individuals without sufficient healthcare access would be more likely to experience frequent poor mental health days, and that models would identify access-related variables as key predictors. The results supported this prediction: access-to-care features were consistently among the most essential variables in the model. This reinforces the conclusion that healthcare accessibility is critical in mental health outcomes and supports further public health efforts to reduce barriers to care.

**Recommendations and Future Work**

**Practical Recommendations:**

**Public Health Policy:**

Healthcare access variables, including insurance status, affordability of care, and access to a primary care provider, were among the top predictors of frequent poor mental health days. This suggests that policies aimed at reducing cost-related barriers, expanding insurance coverage, and improving access to primary care could have a meaningful impact on mental health outcomes.

**Targeted Interventions:**

Public health agencies and community organizations could use similar predictive models to identify high-risk populations and tailor outreach, support services, and preventative programs accordingly. For example, individuals who report difficulty concentrating due to mental health conditions or live with someone who is depressed or suicidal may benefit from targeted mental health screenings and resources.

**Use of BRFSS for Ongoing Monitoring:**

Our project reinforces the BRFSS dataset's value for behavioral health surveillance. Agencies could adopt machine learning tools like Random Forest on BRFSS data for annual mental health trend analysis and to support early identification efforts at scale.

**Future Work and Next Steps:**

If we had more time, we would explore several enhancements to improve our modeling and broaden our insights:

- **Model Expansion:** We would test additional algorithms, including neural networks, to see if predictive performance could improve beyond Random Forest.
- **Probability Modeling:** Rather than using a hard classification threshold, we would consider modeling the probability of reporting poor mental health and calibrating predictions to reflect real-world risk levels better.
- **Class Imbalance Solutions:** We would further explore techniques like SMOTE or ensemble resampling to more robustly address the class imbalance between "Low" and "High" cases.

- **External Validation:** To test model generalizability, we would apply our approach to another year of BRFSS data or a comparable dataset and evaluate performance across different populations.
- **Policy Simulation:** With additional resources, we would simulate the impact of proposed interventions, such as expanding insurance coverage, on predicted mental health outcomes, using counterfactual or what-if analysis techniques.

These next steps could provide deeper insight into the relationship between healthcare access and mental health, and offer a stronger foundation for public health planning and decision-making.