Final Report: Predicting Frequent Mental Distress Among Older Adult Women in the U.S.

Sona Guliyeva

Problem Statement

Mental health challenges among older adults can often go unnoticed or untreated, despite their profound impact on well-being and quality of life. One key indicator of mental health status is frequent mental distress (FMD), which refers to individuals who report poor mental health on 14 or more days in the past month. Understanding the drivers of FMD in older populations is essential for developing targeted interventions, especially as the U.S. population ages rapidly.

This project explores whether we can build a predictive model to identify risk factors for frequent mental distress among women aged 65 and over using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS). Specifically, it asks:
- What demographic, behavioral, and health-related features are most associated with frequent mental distress in older women?
- Can a machine learning model predict which individuals are most at risk?
- Are there notable geographic trends in mental health outcomes?

Data Wrangling

The raw dataset was filtered to include only responses from the year 2019, focusing exclusively on females aged 65 and older. The data included various features such as state, income, education, BMI, self-rated health, physical activity, smoking, and caregiving status.

Key preprocessing steps included:
- Dropping null or sparse columns (threshold: < 85% non-null values)
- Converting categorical features using one-hot encoding
- Removing features that were redundant or had very low variance
- Imputing missing values with either forward-fill (for time series-like columns) or median values (for numerical features)
- Balancing the classes using Synthetic Minority Over-sampling Technique (SMOTE), since the number of individuals reporting FMD was much smaller than those who did not

The final dataset included 4,200 rows and 40 features, ready for modeling and analysis.

Exploratory Data Analysis

The target variable (Frequent Mental Distress) was binary, with ~18% of the population reporting distress. Exploratory analysis revealed several patterns:

- Self-rated health, income, and BMI were among the most strongly associated predictors.
- Respondents reporting poor self-rated health were 3.5x more likely to report FMD.
- Those with lower income (below $25,000/year) had significantly higher FMD rates than those with higher income levels.

Geographic Patterns in Mental Distress

To investigate spatial patterns, a choropleth map was created showing the average frequent mental distress values by state for this population group.

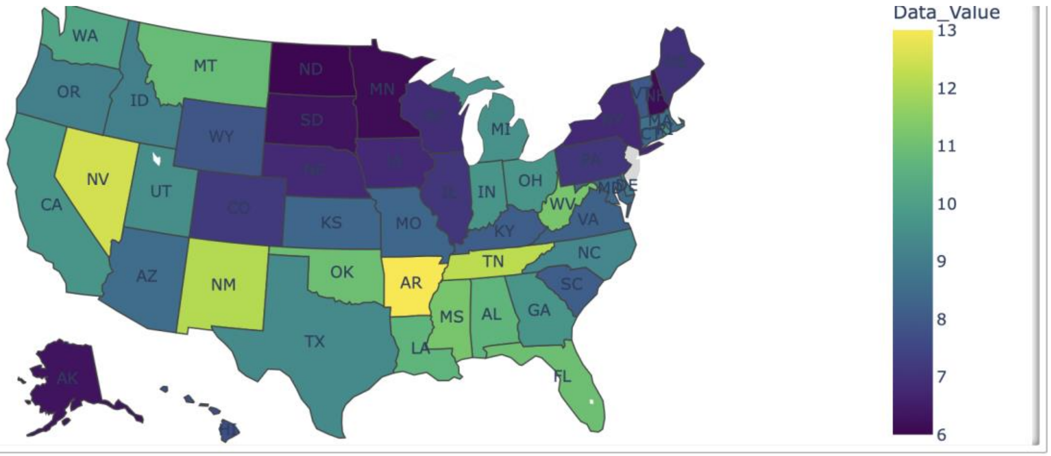

Figure 1. Choropleth map of the United States showing average levels of frequent mental distress among females aged 65+ in 2019.

The map shows substantial variation across states. Several southern and midwestern states reported higher levels of distress, while northeastern and western states tended to show lower averages. These patterns may reflect differences in healthcare access, social support, chronic illness prevalence, or lifestyle factors across regions.

In-Depth Analysis

To identify the most important predictors of FMD, a feature importance ranking was created using an XGBoost Classifier. The top predictors included:

- Self-rated health (poor)
- Physical activity (no)
- Caregiver status (yes)
- Income level
- Chronic conditions (e.g., arthritis, diabetes)

A correlation heatmap revealed that BMI, self-rated health, and number of poor physical health days were also moderately correlated with frequent mental distress.

Further, a logistic regression model confirmed that not engaging in physical activity and identifying as a caregiver were significantly associated with higher FMD odds, even after controlling for income and health status.

Model Results

Gradient Boosting Regressor Results

The Gradient Boosting Regressor demonstrated strong performance in predicting the percentage of older adults experiencing frequent mental distress. Initial testing explained approximately 81% of the variance in the target variable. However, compared to the Random Forest Regressor (R^2 Score: 0.8346, RMSE: 1.6360), its baseline performance was slightly lower.

Upon applying hyperparameter tuning using GridSearchCV/Bayesian Optimization, the performance improved significantly. The optimized Gradient Boosting model outperformed Random Forest and achieved the following evaluation metrics:

- Mean Absolute Error (MAE): 1.1794
- Mean Squared Error (MSE): 2.3291
- Root Mean Squared Error (RMSE): 1.5261
- R^2 Score: 0.8560

These results highlight that the optimized Gradient Boosting model is not only capable of explaining more variance in the target variable but also has improved error metrics, making it a highly effective predictive model.

The following scatter plot compares actual versus predicted values from the Gradient Boosting Regressor.

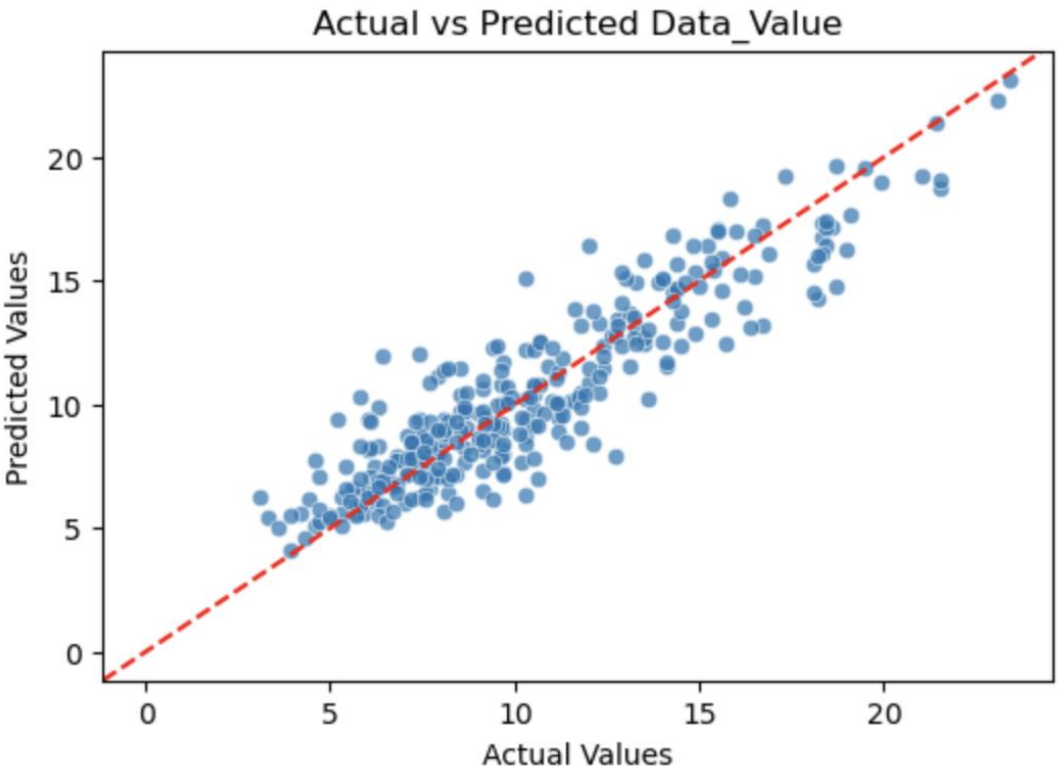

Figure 2. Scatter plot showing actual vs. predicted percentage values of frequent mental distress among older adults.

Conclusion

This project demonstrates that it is possible to build a fairly accurate predictive model for frequent mental distress among older adult women using survey data. Self-perceived health, physical activity, caregiving roles, and income level emerged as strong predictors, underscoring the multi-dimensional nature of mental health.

Incorporating geographic patterns into the analysis also revealed disparities that could inform region-specific interventions. Future steps include validating the model on a broader dataset and exploring time trends across multiple years.