# Week 07: Final Project Milestone (Section 2, Team 4)

Andres Salcedo - asalcedo@berkeley.edu, Cassidy Bruner - cassidybruner@berkeley.edu,
Eyup Agtepe - eyupagtepe@berkeley.edu, Gabrielle Doran - gabrielle_doran@berkeley.edu
**GitHub Repository**: https://github.com/cassidy-bruner/mids-207-final-project-team-4.git

**Motivation:** Alcohol misuse is a major public health concern that contributes to significant physical, psychological, and social harm. In the United States, excessive alcohol consumption remains one of the leading causes of preventable disease and death. Prior research, such as the study by Felitti et al. (1998) in the *American Journal of Preventive Medicine*, revealed a strong graded relationship between Adverse Childhood Experiences (ACEs) and the likelihood of developing alcoholism and other risky behaviors in adulthood. These findings highlight the long-term behavioral consequences of early-life trauma. Our team is motivated to explore whether machine learning can uncover and quantify similar patterns using recent survey data. Specifically, we aim to analyze which demographic and health factors predict risky drinking, and to what extent ACE factors contribute beyond those variables.

**Data + Preprocessing:** We used data from the CDC's 2024 Behavioral Risk Factor Surveillance System (BRFSS), a nationally representative survey of U.S. adults. The survey is administered via telephone interviews and collects information on a wide range of health-related risk behaviors, chronic health conditions, and preventive service use.

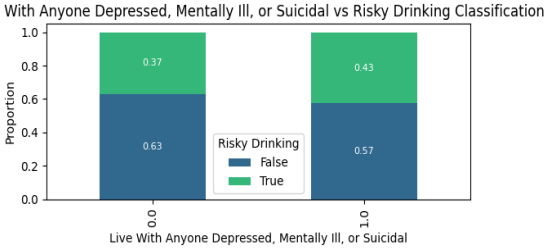
The BRFSS 2024 dataset contains 457,670 respondents across 301 variables. We filtered to 39,283 respondents from 8 states where the ACE module was administered. We engineered a broader risky drinking target combining binge drinking (4+ drinks for women or 5+ drinks for men on one occasion), heavy drinking (8+ drinks for women or 15+ drinks for men per week) and high-frequency drinking (4+ drink occasions per day). This expanded prevalence from 12% to 37% of respondents, creating a more balanced classification problem. We tested this definition on a random 20% sample of all respondents and confirmed similar prevalence (37.4%).

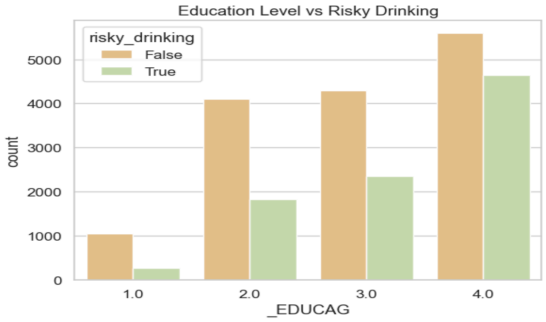
We replaced BRFSS missing response codes 7, 77, and 777 (Don't know) and 9, 99, 999 (Refused) with NaN values. For MENTHLTH and PHYSHLTH we recoded 88 (None) to 0 days. We examined continuous variables for outliers using the IQR method with 3 times IQR as the threshold and retained all values as flagged cases were valid responses of 30 poor health days. We recoded all categorical variables to numeric formats appropriate for modeling. For binary ACE variables we recoded (Yes/No) 1/0, ordinal ACE variables (Never/Once/More than once) to 0/1/ and protective ACE factors (1-5 scale) to 4-0 where higher values indicate less protection. We recoded sex to 0/1 and applied one-hot encoding to the imputed race variable, creating 5 binary indicators with White as reference. We created two ACE features: at_least_1_ace (binary indicator) and ACE_SCORE (cumulative count, range 0-17, mean 1.83). After removing missing values we retained 34,513 complete observations. We selected 25 features: 12 individual ACE variables, 6 demographic and health variables (education, age, sex, mental health days, physical health days, smoking status), 5 race indicators, at_least_1_ace and ACE_SCORE. We excluded income and marijuana use due to high missingness rates. We split the data using stratified sampling with shuffling and random_state set to 42 for reproducibility. The training set contains 24,158 observations (70%), validation set contains 5,178 observations (15%) and test set contains 5,177 observations (15%). We applied StandardScaler to normalize all features, fitting the scaler on training data only and transforming all three sets to prevent data leakage.

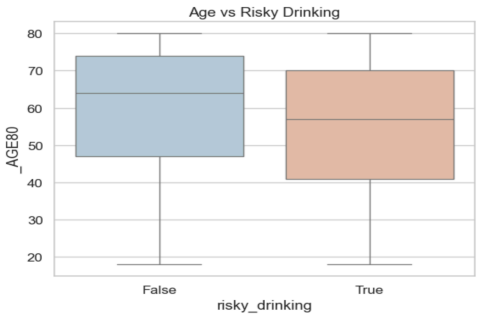
**Data Challenges/Limitations:** Limited coverage of key variables: Only 8 states (Florida, Georgia, Hawaii, Virginia, North Dakota, Nevada, Puerto Rico, the Virgin Islands) included ACE data. These variables are important to our motivation, so we restricted the dataset to these states, reducing sample size, geographic diversity and the model's generalizability. Outcome variable construction: The target variable, risky_drinking, was derived by combining three BRFSS questions on binge, frequent, and heavy drinking. Respondents were labeled as risky drinkers if they met our specified thresholds in any of the questions. This approach captures multiple risk types but may introduce bias from the chosen thresholds. Categorical and complex variable structure: Most BRFSS features are categorical and numerically coded, requiring cross-referencing documentation sources to interpret accurately. This process was time-consuming, and preparing the data for modeling demanded extensive encoding and cleaning to address missing or ambiguous responses.

**Exploratory Data Analysis:** We performed bivariate analyses to explore associations between demographic and behavioral features and target: risky_drinking. The plot left and below shows the proportion of individuals classified as engaging in risky drinking (green) versus non-risky drinking (blue) by sex. Among men (0), 44% are classified as risky drinkers. Among women (1), 32% are risky drinkers. This suggests, men are more likely to engage in risky drinking than

women in this sample. The plot to the right shows the proportion of individuals classified as engaging in risky drinking (green) versus non- risky drinking (blue) by the response to “Live With Anyone Depressed, Mentally Ill, Or Suicidal?”. Among those who said no (0), 37% are classified as risky drinkers. For those who said yes (1), 43% are risky drinkers.

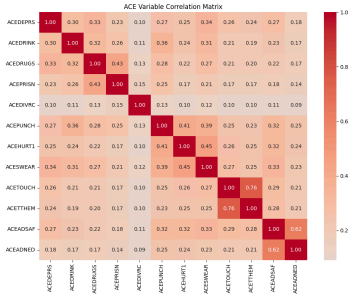



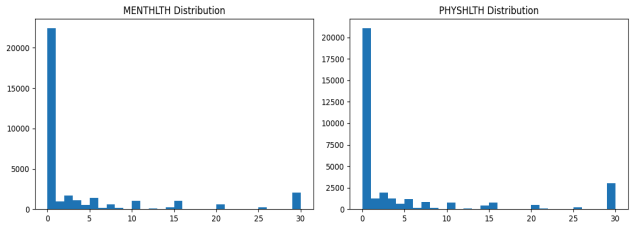




In the Age vs Risky Drinking boxplot (right above), the median age of risky drinkers appears noticeably younger, while non-risky drinkers are generally older. The distributions show that younger adults are more likely to engage in risky drinking behaviors, while older adults tend to report lower rates. Both groups have a few respondents in the lower age range, but there are no significant outliers.

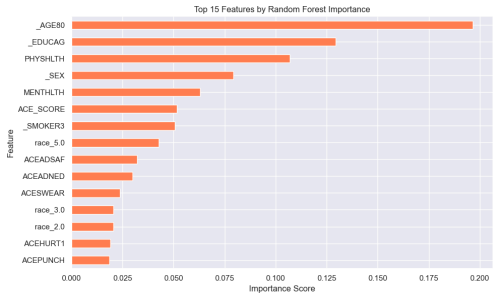
In the Education Level vs Risky Drinking bar chart (left above), individuals with higher education levels (categories 3 and 4, attended or graduated college) report more instances of risky drinking compared to those with lower education. However, the overall number of non-risky drinkers remains slightly higher in every education group. This pattern may indicate that people with higher education levels are more likely to engage in social drinking.




We also examined correlations among the Adverse Childhood Experiences (ACE) variables to identify potential multicollinearity and shared patterns of adversity. As shown in the correlation heatmap on the left, most ACE variables exhibit moderate positive correlations (r = 0.2–0.4), suggesting that individuals reporting one type of adverse experience often report others.




We examined the distributions of the variables. You can see to the right, both *MENTHLTH* and *PHYSHLTH* variables are right-skewed, with most respondents reporting 0 unhealthy days and smaller groups reporting more. This suggests that while most individuals in the sample report good overall health, a small subset experiences chronic mental or physical health difficulties, which could be potentially relevant factors in understanding risky drinking behavior.




We assessed featured importance using two methods. Pearson correlation coefficients were calculated between each feature and the target variable to measure linear relationships. A Random Forest classifier was trained using 100 trees and max depth 10 to extract feature importances. This captured non-linear relationships and interactions.

Demographic and health variables were the strongest predictors of risky drinking. Education level showed the highest correlation (r=0.150). Age had the highest Random Forest importance (0.197). The top five predictors by both methods were education, age, sex, physical health days and race.

Among ACE variables, ACESWEAR (verbal abuse) had the strongest correlation (r=0.060). The cumulative ACE_SCORE ranked 6th overall in Random Forest Importance (0.052). The binary at_least_1_ace indicator showed weak predictive value with the second to lowest rank for Random Forest importance. Individual ACE experiences showed correlation ranging from r=0.002 to r=0.060. Figure x gives us a better summary of the top 15 features ranked by importance.

**Methodology:** We will employ a progressive modeling strategy to predict risky drinking behavior using the features described above. This strategy will follow the machine learning pipeline methodologies depicted so far in 207, including:

1. **Baseline Model - Majority Class Classifier:** Our baseline will predict the majority class for all observations (No risky drinking), providing a simple benchmark for comparison. This establishes the minimum threshold that a more sophisticated model must improve upon to be useful.
2. **Logistic Regression:** Logistic regression will be our first improvement over baseline. This model is well-suited for binary classification problems (risky drinking = 1, no risky drinking = 0). Logistic regression outputs probability estimates that are useful for risk assessment, and feature weights provide interpretable coefficients that quantify how ACEs and other features contribute to risky drinking behavior.
3. **Neural Network:** Currently, a feedforward neural network will be our most complex model, but as we learn new techniques this semester, we will implement them into our methodology. Based on implementation using keras.Sequential(), our architecture would include an input layer; multiple hidden layers with ReLU or tanh activations for non-linear pattern recognition; an output layer with softmax activation for probability estimation; and an SGD optimizer with tunable learning rates.

For hyperparameter tuning and validation, we will use Keras Tuner to optimize the neural network, testing different configurations of hidden layers, activation functions (ReLU, tanh), learning rates (0.001, 0.01, 0.1), batch sizes, and dropout rates. For logistic regression, we will tune L1/L2 regularization parameters to prevent overfitting. We will evaluate models using accuracy, precision, recall, and F1 score.

In evaluating our models, we will consider two perspectives: a cost-saving perspective, and a risk mitigation perspective. In the cost-saving perspective, if resources are limited, we prioritize precision to minimize false positives. Over-classifying individuals as risky drinkers wastes intervention resources (counseling, follow-ups) that could be directed to those who truly need them. For the risk mitigation perspective, from a public health standpoint, we would prioritize recall to minimize false negatives. Missing someone with risky drinking behavior means missing opportunities for early intervention that could prevent alcohol use disorders or other consequences. We will present results from both perspectives using ROC and precision-recall curves, allowing stakeholders to make informed decisions based on their priorities and resource constraints.

**Contributions**

Cassidy created/organized the team's GitHub repo, hosted Zoom meetings and created the outline of all our google docs. She reviewed the preprocessing notebooks (added code to export cleaned datasets and comments/documentation). She authored cassidy_eda.ipynb and conducted EDA focusing on the correlation of features with the target. She came up with models, evaluation metrics, and two perspectives for the methodology. She wrote the data challenges/limitations section.

Eyup performed independent data preprocessing but did not include it in the GitHub repository to avoid confusion. He conducted exploratory data analysis focusing on the relationship between Adverse Childhood Experiences and risky drinking behavior. His analysis compared ACE vs. non-ACE groups, considering factors such as age and education level.

Gabby performed feature engineering and testing of the target class by incorporating drinks per day into its calculation. She contributed to EDA, with a focus on investigating relationships between race, ACE counts, and drinking behavior. She wrote the Methodology section of this submission, depicting the ML techniques to be utilized in this project.

Andres conducted data preprocessing including missing value handling, feature engineering (ACE_SCORE, risky_drinking), categorical encoding and train/val/test splitting. Performed EDA while focusing on feature importance analysis using correlations and Random Forest methods. He wrote the data preprocessing and feature importance analysis parts. He created the BRFSS_data_preprocessing and andres_eda notebooks.