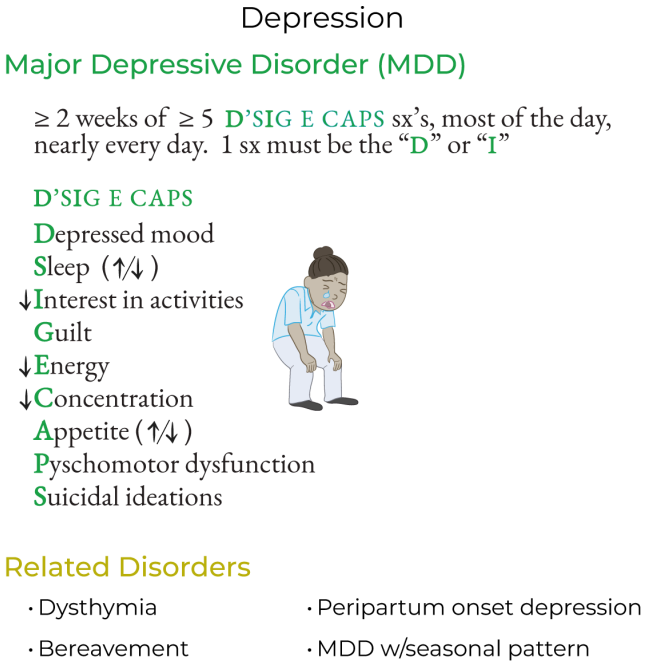
Predicting Major Depressive Episodes in US
MadHacks 2026 - Reflect Team

RESEARCH QUESTION

This analysis asks a different question: *can we predict MDE risk from information we already know about people, their age, sex, marital status, employment, substance use patterns, without requiring any clinical screening?* If the answer is yes, that opens the door to population-level early intervention.

PURPOSE

This analysis uses NSDUH survey data from 180,034 U.S. adults to identify who is most at risk for Major Depressive Episode (MDE). Using logistic regression and random forest classification, we find that marijuana use frequency and marital status are the strongest demographic predictors of MDE, outperforming poverty, employment, and insurance coverage. A 0.722 AUC random forest model built entirely from demographics and behavior (no symptom measures) demonstrates that at-risk individuals can be meaningfully identified before clinical screening. These findings have direct implications for targeted mental health outreach and resource allocation.



BACKGROUND

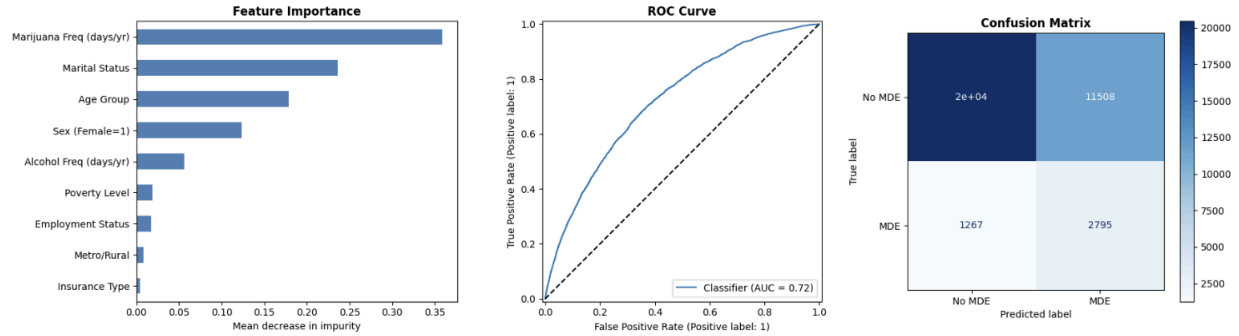
Survey symptoms from NSDUH which fulfill the conditions for Major Depressive Episode (MDE) affect 1 in 9 US adults annually, yet the majority of people never receive treatment or formal diagnosis. Identifying who is at risk before clinical diagnosis could enable earlier intervention.

METHODS

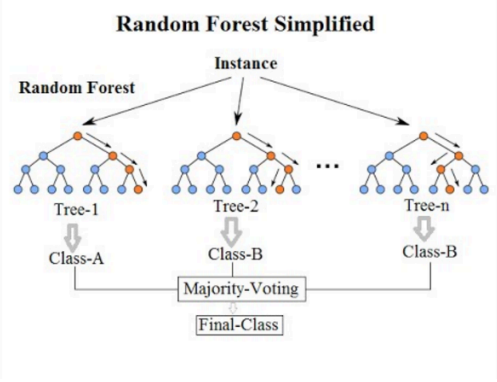
We used the NSDUH 2021–2024 combined public-use file, restricting to adults 18+ (n=180,034). We recoded all binary variables to 0/1 and dropped null values We utilized NSDUH survey data from 180,034 U.S. adults (18+) to identify who is most at risk for Major Depressive Episode (MDE). Selected variables included MDE diagnosis in the past 12 months, Age Group, Poverty Level, Marital Status, Employment Status, Days Drinking in the past year, Days using Marijuana in the past year, Insurance Type, and rural/urban.

RESULTS

We used a Random Forest Model with 300 trees with max depth 6, class_weight='balanced' to address the 11.3% MDE prevalence (class imbalance), and an 80/20 train-test split stratified by outcome.



The model achieved **0.722 AUC** on the held-out test set (n=36,007). Of 4,062 adults with MDE in the test set, 2,795 were correctly identified (68.8% recall). The top predictors were marijuana use frequency (35.8%), marital status (23.6%), age group (17.8%), and sex (12.3%). Poverty, employment, insurance, and geography together contributed less than 5% — suggesting their effect on MDE is largely mediated through substance use and relationship stability rather than operating independently.



LIMITATIONS

NSDUH surveys different people each year. We cannot establish causality or track individuals over time. The marijuana-MDE relationship could run in either direction.

In addition, we were not able to analyze by state. The public-use file suppresses state identifiers for privacy. Our geographic analysis is limited to metro vs. rural. We cannot identify regional variation. Our prediction variable is affected by class-imbalance. MDE affects 11.3% of the sample. Despite class_weight='balanced', the model still produces many false positives (11,508 in test set). Furthermore, all variables including MDE are based on self-report. Stigma around mental health may lead to underreporting, particularly in certain demographic groups. Four years of data limits trend analysis. Extending to pre-2015 data requires careful harmonization due to NSDUH methodology change.

NEXT STEPS

Spearman correlation heatmap matrix across all 10 variables to understand co-occurrence patterns between MDE, suicidality, substance use, and demographics at the individual level. Cross validation to evaluate different models and parameters. In particular, it could be interesting to investigate gradient boosting compared to random forest. A direct comparison would establish the ceiling of predictive performance for this feature set.

WORK CITED

SAMHSA. (2023). *Key substance use and mental health indicators in the United States: Results from the 2022 National Survey on Drug Use and Health*. HHS Publication No. PEP23-07-01-006. https://www.samhsa.gov/data/

AMHSA. (2024). *National Survey on Drug Use and Health (NSDUH), 2021-2024* [Data file and codebook]. Substance Abuse and Mental Health Services Administration. https://www.datafiles.samhsa.gov