THE SCHOOL OF MATHEMATICS AND STATISTICS

WELLINGTON FACULTY OF ENGINEERING

DATA301- DATA SCIENCE IN PRACTICE

INDIVIDUAL PROJECT REPORT

DEPRESSION PREDICTION USING DEMOGRAPHICS AND DIETARY NHANES DATA

Group 11 Project Members:

	Name	**Student ID**
1	Juliet Jefferis	300563977
2	Aidan Malcolm	300584080
Individual	Lam Quang Thinh	300538520

Executive Summary

The NHANES dataset, sourced from the National Health and Nutrition Examination Survey, consists of detailed records from three primary sections: dietary, demographic, and questionnaire (with emphasis on the depression label). The dataset captures a vast array of health metrics, dietary profiles, and demographic details of participants. Given the health and personal nature of the NHANES dataset, potential biases might be present, stemming from the data collection and representation processes. This could affect the representation of specific demographic groups or health conditions. There are minimal privacy concerns due to the anonymized nature of the data, ensuring participant confidentiality.

The dietary, demographic, and questionnaire datasets were seamlessly integrated using respondents sequence number (SEQN) identifiers, ensuring consistency across records. During this merging process, challenges such as handling missing values and ensuring uniformity among categorical variables were addressed.

The primary objective of the project was to *unearth patterns and correlations between various factors in the dataset and the depression status of participants*. Additionally, a predictive model was developed to estimate the likelihood of depression based on various factors, shedding light on potential risk elements. In specific, initial investigations have pointed towards potential relationships between certain dietary components and depression status. Moreover, demographic factors like gender, origin, and education level appeared to have significant associations with depression. Beyond the preliminary analyses, a predictive model was developed to *predict depression based on other factors in the datasets*. This model not only helps in understanding the intricate relationships between variables but also serves as a tool for early identification of at-risk individuals.

The NHANES dataset is a treasure trove of information that can be harnessed to understand the multifaceted nature of depression. Through rigorous analysis and predictive modeling, we have the potential to derive actionable insights, facilitating informed interventions and policy-making. However, it's essential to approach these findings with caution, always being mindful of the potential biases and limitations inherent in the data.

Table of Contents

I. Background

The dataset employed for our research project is sourced from the National Health and Nutrition Examination Survey (NHANES). NHANES is a prominent program of studies, meticulously designed to evaluate the health and nutritional status of adults and children across the United States. The survey is orchestrated by the National Centre for Health Statistics (NCHS), a vital arm of the Centers for Disease Control and Prevention (CDC). Since its inception in the 1960s, NHANES has undergone various evolutions, with diverse survey objectives over the years. From 1999 onwards, it adopted a consistent format, surveying approximately 5,000 American civilians annually, excluding those institutionalized.

Our research focuses on NHANES data from the period of 2013 to 2020. This rich trove of information is publicly accessible on the official website of the Centers for Disease Control and Prevention. Historically, NHANES datasets have been instrumental in myriad research initiatives, ranging from identifying potential associations between chemical exposures and chronic diseases to intricate analyses of oral microbiomes. The versatility of the data stems from its comprehensive nature, encapsulating a vast array of predictor variables and diverse medical conditions. Such versatility makes it a prime candidate for studies like ours, aiming to unearth potential correlations between dietary, lifestyle, and environmental factors with the prevalence of depression.

There are two main objectives for our project. Primarily, we aim to unravel the intricate associations between a multitude of factors and depression status. By elucidating these correlations, we strive to enhance our comprehension of potential antecedents or contributors to depression. Building upon this foundational understanding, the secondary objective is to craft a sophisticated machine learning model. This model is envisioned to predict depression status based on the identified factors, thus serving as an invaluable tool for early detection. By facilitating timely prevention and interventions, we aspire to contribute meaningfully to the domain of mental health.

The structural organization of the NHANES dataset is meticulous. Data is methodically segmented into datasets, such as Demographics, Dietary Data, Examination Data, Laboratory Data, and Questionnaire Data, among others. Each category boasts a myriad of datasets, segregated based on biennial periods starting from 1999-2000. These datasets are intricately linked using a unique identifier: the respondent's sequence number (SEQN). Given the vast expanse of available data, our approach was judiciously selective. We amalgamated select datasets from various years, cognizant of a notable transition in variable naming conventions post-2003. One significant challenge we navigated was the absence of direct indicators for depression diagnoses in the data. To surmount this, we resorted to labeling the data based on a subset of NHANES survey questions. Our labeling methodology, encompassing both binary and multi-class categorizations, was informed by these two esteemed sources: NHANES Component Description (2015) and The PHQ-9: validity of a brief depression severity measure (2001).

II. Data Description

The data combined three datasets from the NHANES collection in the US, including Demographics, Dietary, and Questionnaire datasets. After being merged and cleaned, the dataset has 190 variables. The full model were applied to the Catboost algorithm to identify the most important features for the depression prediction. The most influential features derived from the model is selected and represented in Figure 1:

Variable	**Data type**	**Description**
Label	integer	Depression label, has depression (1) or not (0)
yesterday_food_amount	float	Compare food consumed yesterday to usual
moisture	float	Moisture (gm)
total_sugar	float	Total sugars (gm)
dodecanoic_amount	float	SFA 12:0 (Dodecanoic) (gm)
day_gap	float	Number of days between the intake day and the interview
beta_carotene	float	Beta-carotene (mcg)
income_poverty_ratio	float	A ratio of family income to poverty guidelines
age	integer	Age in years of the participant at the time of screening
origin	integer	Country of origin of the participant
interpreter	integer	Was an interpreter used to conduct the Family interview
gender	integer	Gender of the participant
education	integer	Education level

Figure 1 - Important features description

There are 4 variables with missing values: income_poverty_ratio with 8.35%, ‘education’ with 5.43%, interpreter and day_gap with 1.96% and 1.90%, respectively. All missing values in the datasets will contain ‘.’ value.

In specific, there are categorical variables that have encoded value:

- The *‘origin’* variable has four categories: ‘1’ for ‘born in the US’, ‘2’ for ‘Others’, ‘77’ for ‘refused to answer’, ‘99’ for ‘do not know’.
- The *‘Interpreter’* variable has two categories: ‘1’ for ‘yes’, ‘2’ for ‘no’.
- The *‘gender’* variable has two categories: ‘1’ for ‘Male’, ‘2’ for ‘Female’.
- The *‘education’* variable has seven categories: ‘1’ for ‘Less than 9th grade’, ‘2’ for ‘9-11th grade (Includes 12th grade with no diploma)‘, ‘3’ for ‘High school graduate/GED or equivalent’, ‘4’ for ‘Some college or AA degree’, ‘5’ for ‘College graduate or above’, ‘7’ for ‘refused to answer’, and ‘9’ for ‘do not know’.

III. Ethics, Privacy, and Security

Data Collection and Ethical Considerations:

The ethical integrity of our research is rooted in the principle of informed consent. Our chosen dataset, NHANES, exemplifies rigorous data collection standards, having been sourced from American civilians who were comprehensively briefed on both the nature of their contributions and the associated potential risks. Notably, NHANES receives annual validation from an established ethics board. This adherence to stringent ethical protocols ensures that our project remains within the confines of the data's original intent, enabling us to undertake robust health-centric statistical evaluations.

Research Framework and Participant Privacy:

Our research philosophy aligns with the perspective articulated by Prof. Derick Wade, emphasizing that the collection and utilization of patient data must extend beyond immediate clinical decisions to address overarching societal queries. The depth and focus of the NHANES dataset, dedicated to the health assessment of U.S. citizens, serve as pivotal assets in this endeavor. As we synthesize insights from this expansive data, maintaining participant confidentiality remains paramount. We are steadfast in our commitment to ensuring that our disseminated results, while informative, do not compromise individual identities. By adopting NHANES' rigorous privacy protocols, which include encrypted data storage and meticulous data review processes, we fortify our commitment to data confidentiality.

Analytical Fairness and Data Security Protocols:

A cardinal aspect of our research methodology is ensuring representational fairness in our analytical outcomes. NHANES' structured and unbiased data collection mechanisms equip us with a foundation to derive inclusive insights. However, the inherent U.S.-focused orientation of the dataset necessitates circumspection in extrapolating findings to global contexts. To further our commitment to unbiased research, we have incorporated fairness metrics, striving for equitable prediction outcomes across diverse demographic categories. In terms of data security, we have instituted rigorous safeguards. All data is securely housed on designated personal computers, precluding external access. While our methodological blueprint is transparently archived on GitHub for academic scrutiny, we have ensured that raw data remains inaccessible. Additionally, we are exploring advanced cryptographic solutions to enhance our data protection measures further.

IV. Exploratory Data Analysis

The distribution in Figure 2 reveals that a larger portion of the dataset consists of individuals without depression, compared to those with depression. This indicates an imbalance in the dataset, which is a common scenario in medical or health-related data.

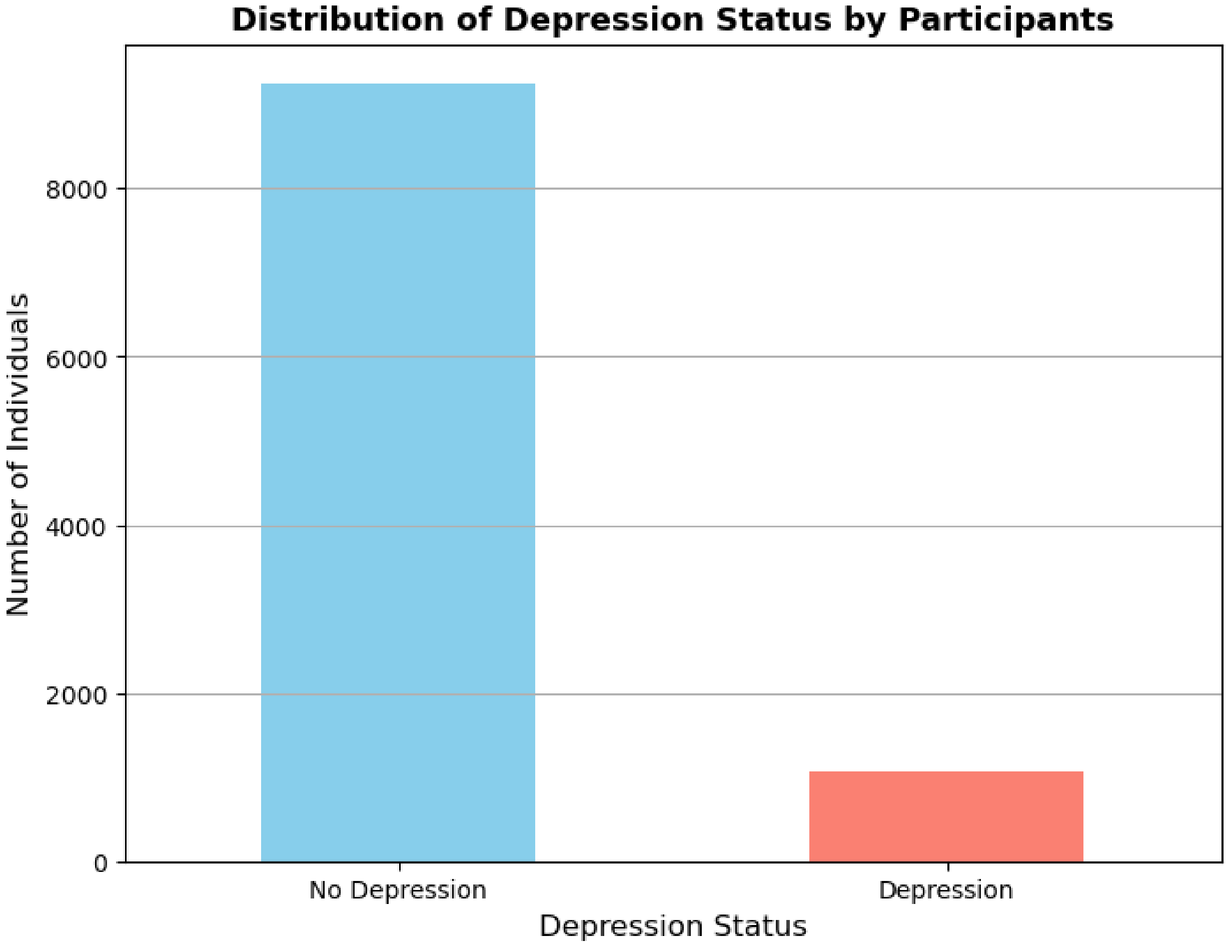

Figure 2 - The distribution of Depression status of participants

Figure 3 illustrates the Distribution of Continuous Variables by Depression Status, which provides an overall distribution between participant with and without depression diseases:

- *yesterday_food_amount*: There does not seem to be a significant difference in the amount of food consumed the previous day between those with and without depression.
- *moisture*: There's a slight difference in moisture levels, with depressed individuals showing a slightly lower median moisture level compared to non-depressed individuals.
- *total_sugar*: There is no pronounced difference in total sugar intake between the two groups.
- *dodecanoic_amount*: Individuals with depression tend to have a slightly higher median intake of dodecanoic acid than those without depression.
- *beta_carotene*: The median amount of beta-carotene seems to be lower for individuals with depression compared to those without.

- *income_poverty_ratio*: There is an indication that individuals with depression might have a decent lower median income-poverty ratio.

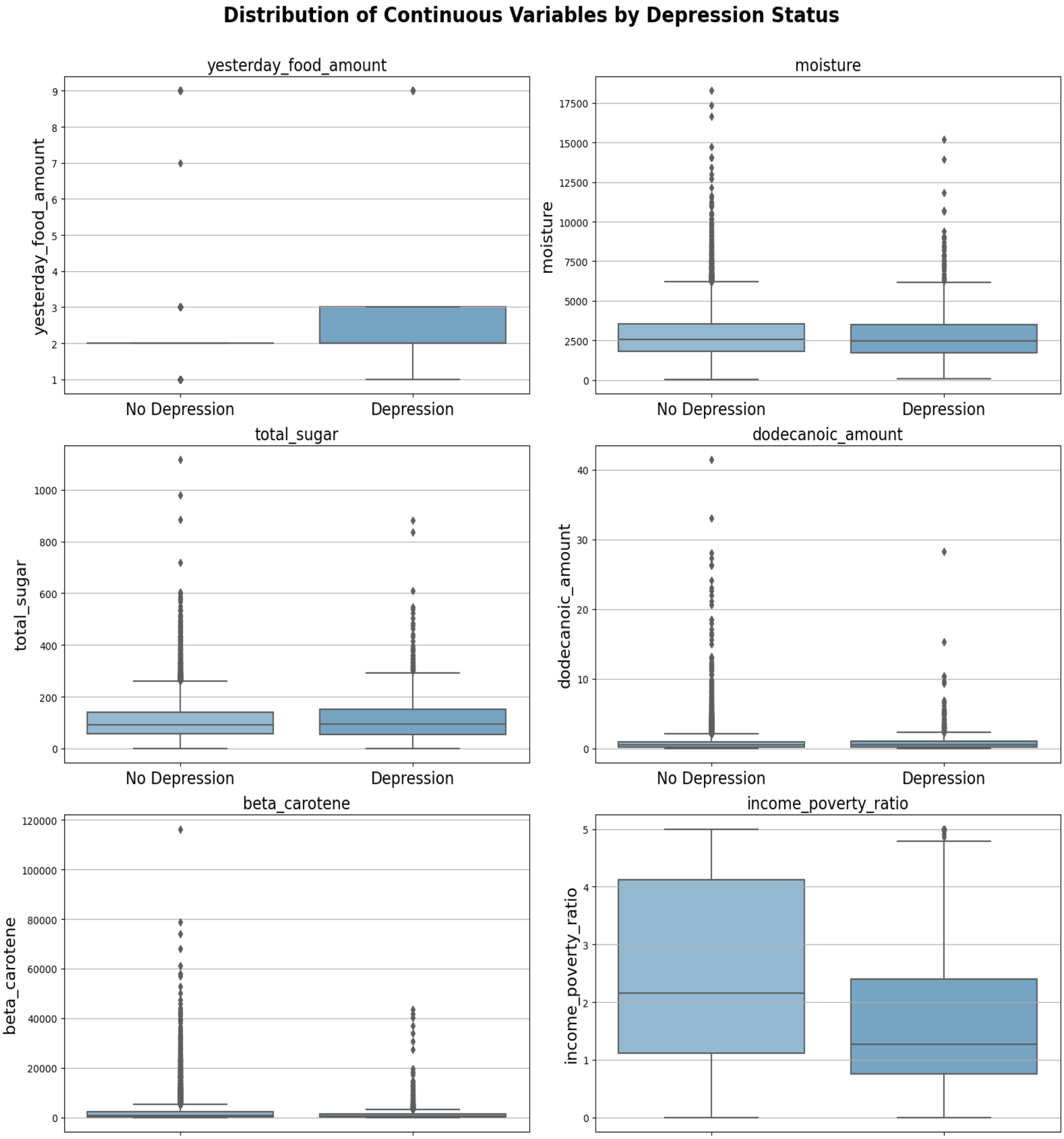

Figure 3 - Distribution of Continuous Variables by Depression Status

The bar plots in Figure 4 provide insights into the distribution of various categorical variables based on depression status. It is worth noting that the sample size for some categories might be small, so these findings should be interpreted with caution.

- *Origin*: The proportion of individuals with depression seems to vary based on the origin category. The participants that refused to answer this question seem to have the highest possibility to have depression.
- *Interpreter*: The proportion of individuals with depression appears to be higher for those who did not have an interpreter for the interview, compared to those who have.
- *Gender*: The proportion of individuals with depression seems to be higher for Female compared to Male.
- *Education*: The proportion of individuals with depression varies based on education levels. Unknown education level have the highest proportion of individuals with depression

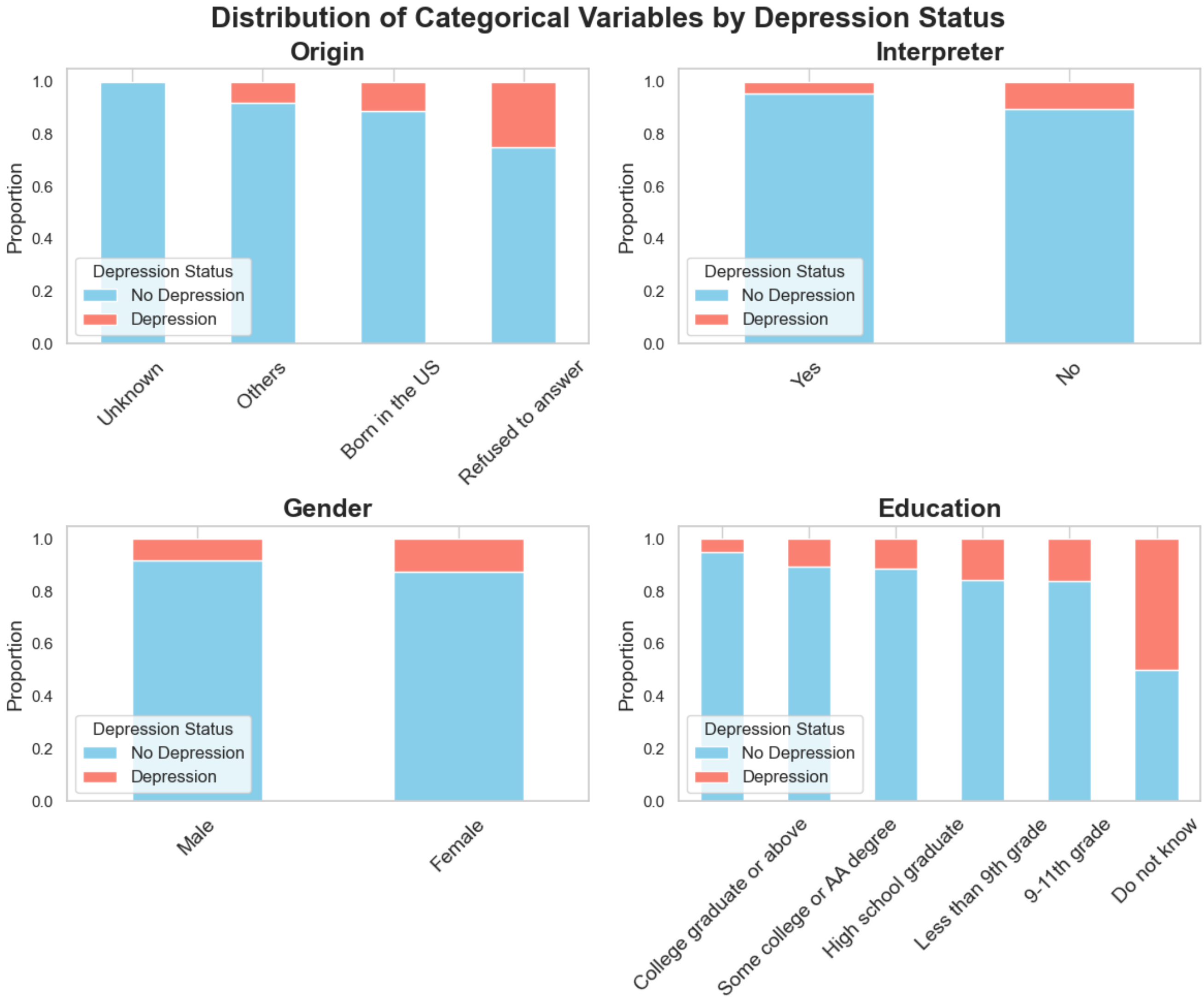

Figure 4 - Distribution of Categorical Variables by Depression Status

V. Detailed Analysis Results

1. Statistical Analysis

After having an overall on the features as individuals, it is essential to observe their associations to each other and to the target variable.

1.1 Continuous variables

The correlation heatmap in Figure 5 shows the pairwise correlations between continuous variables. In specific, all correlations are relatively low, indicating that there is almost no strong linear relationship between the features. Some pairs, like moisture and total_sugar, dedecanoic_amount and total_sugar, show mild positive correlations. Therefore, there is no concern for collinearity among these features

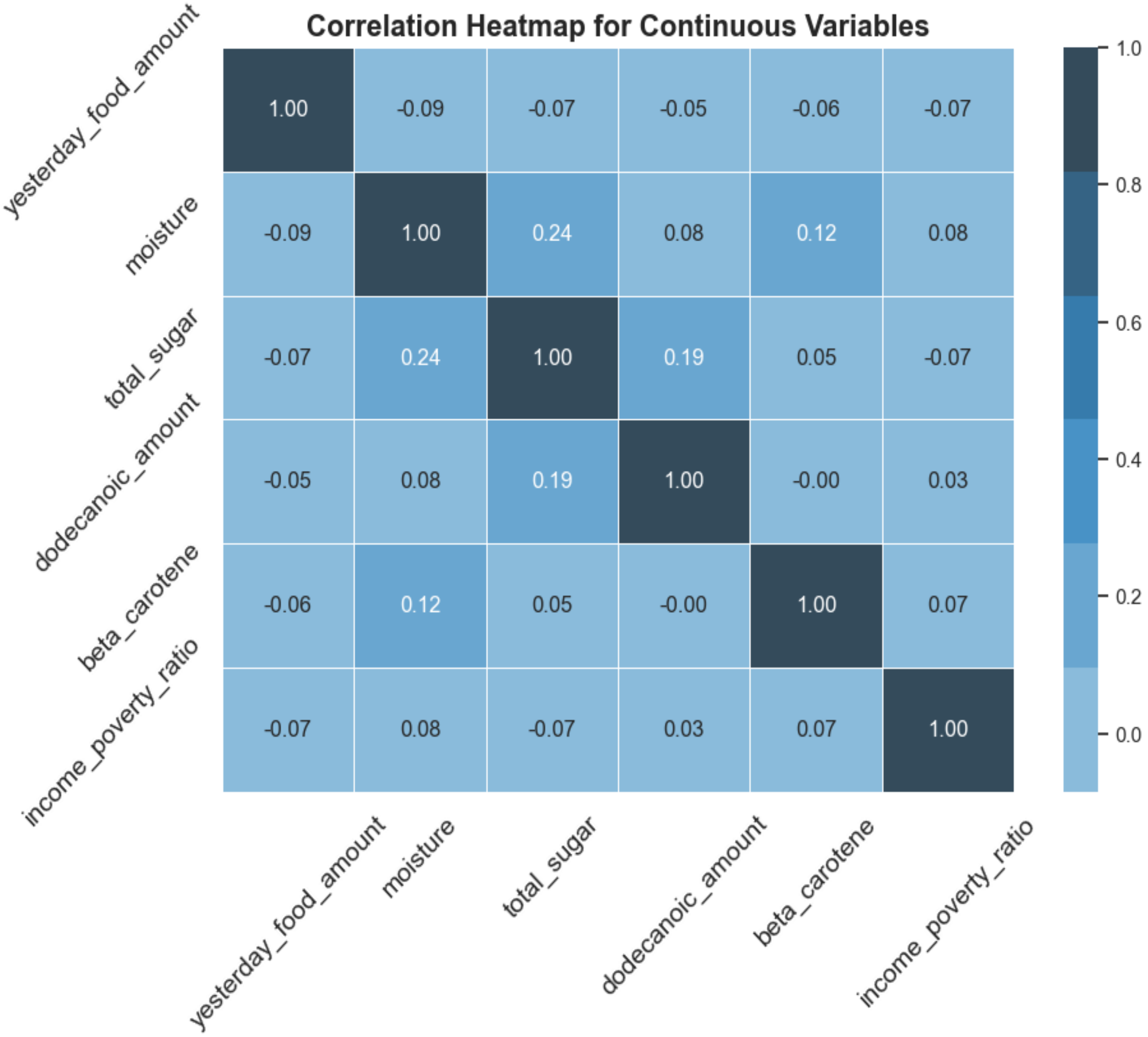

Figure 5 - Correlation Heatmap for Continuous Variables

1.2 Categorical variables

Figure 6 visualizes the significance of the Chi-squared test results for the categorical variables, the higher the bar, the more significant the association between the categorical variable and the Label column. Also, the red dashed line represents the significance level at α=0.05 and any bar above this line indicates a significant association. From the plot, all categorical variables have a significant association with the Label column since their bars are above the red dashed line. Specifically, the education variable has the highest significance, followed by gender, origin, and interpreter. Consequently, these variables should be included to observe their predictive powers.

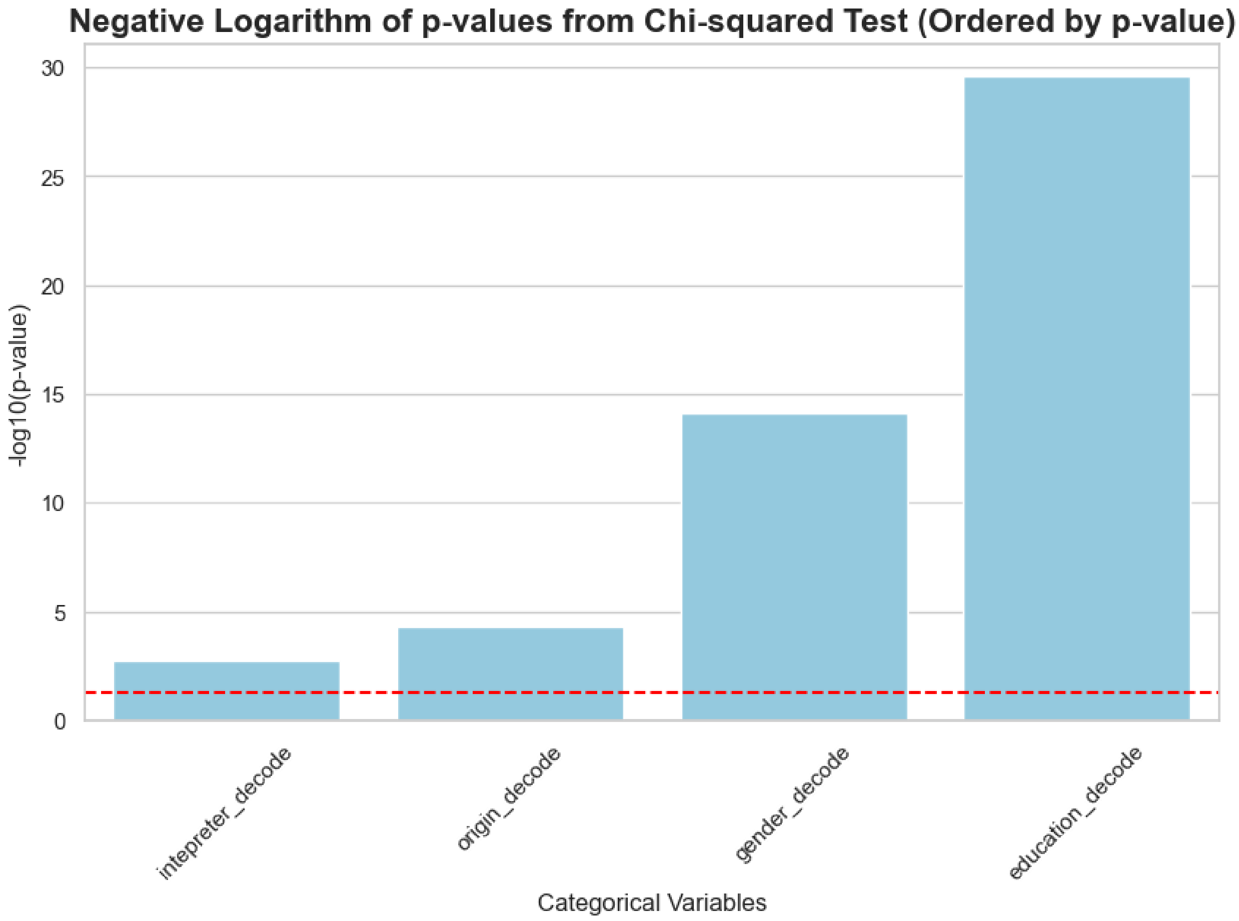

Figure 6 - Chi-squared Test for Categorical Variables

2. Machine learning predictive models

The model development process includes the following steps:

2.1. Data preprocessing:

- Split the data into training and test sets with the ratio 80:20
- Impute the missing values for numerical features by mean
- Standardize the numerical features by using StandardScaler
- Impute missing values for categorical features by mode
- Encode the categorical features

- Create a pipeline for all of the steps above
- Fit and transform the training set
- Transform the test set

2.2 Model training

- Initialize some machine learning algorithm classifiers: Catboost, XGBoost, and LightGBM
- Train the classifiers using the training data

2.3 Initial model evaluation

- Predict on the test data and observe the results as below:

Model	Accuracy	Precision	Recall	F1 Score	ROC- AUC
Catboost	0.9007	0.5000	0.0146	0.0284	0.6661
XGBoost	0.8959	0.3529	0.0584	0.1004	0.6389
LightGBM	0.8978	0.2857	0.0195	0.0365	0.6614

Figure 7 - Model comparisons

According to the model comparisons in Figure 7, Catboost tends to have the best performance across the metrics. Therefore, it is selected to perform the next step: hyper-parameter tuning.

2.4 Hyperparameter tuning for Catboost model

Model	Accuracy	Precision	Recall	F1 Score	ROC- AUC
Tuned Catboost	0.9007	0.5000	0.0146	0.0284	0.6661

Figure 8 - Tuned Catboost model

After tuning, the optimal hyperparameters for the Catboost model are the default hyperparameters. Therefore, the performance of the model remains the same as Figure 8.

2.5 Final model evaluation

After finalizing the best model, there are comments on the model results. In detail, the high accuracy alongside a low recall suggests that the model is likely predicting the majority class (no depression) most of the time. The ROC-AUC score, being above 0.5, indicates that the model performs better than random guessing, but there is a significant margin for improvement. Finally, the most important metric, the low recall, indicates that the model is not capturing most of the positive cases (individuals with depression). This can be problematic, especially because the goal is early detection and intervention for those at risk of depression.

2.6 Model interpretation

According to the Feature Importance (Figure 9) and the SHAP values (Figure 10), the `*income_poverty_ratio*` feature emerges as a dominant predictor. The `*dodecanoic_amount*` variable also holds notable significance, warranting further investigation into its specific role. Age consistently stands out in its influence, indicating possible variations across different age segments. The `*education*` feature indicates a distinct impact, suggesting potential multifaceted interactions. Dietary attributes, including `*yesterday_food_amount*`, `*dodecanoic_amount*`, and `*beta_carotene*`, manifest as influential factors. The `*gender*` feature also presents itself as a distinguishing factor, alluding to different patterns between the categories.

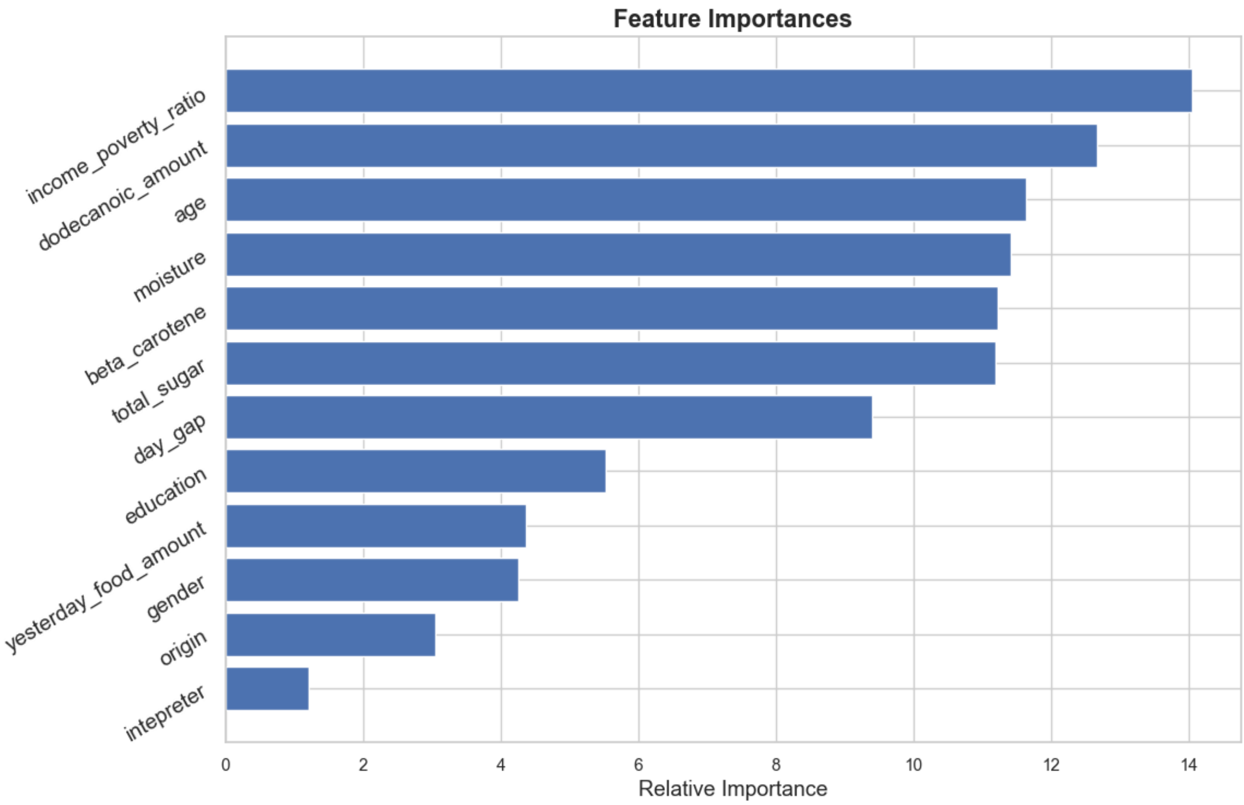

Figure 9 - Feature importance of the final Catboost model

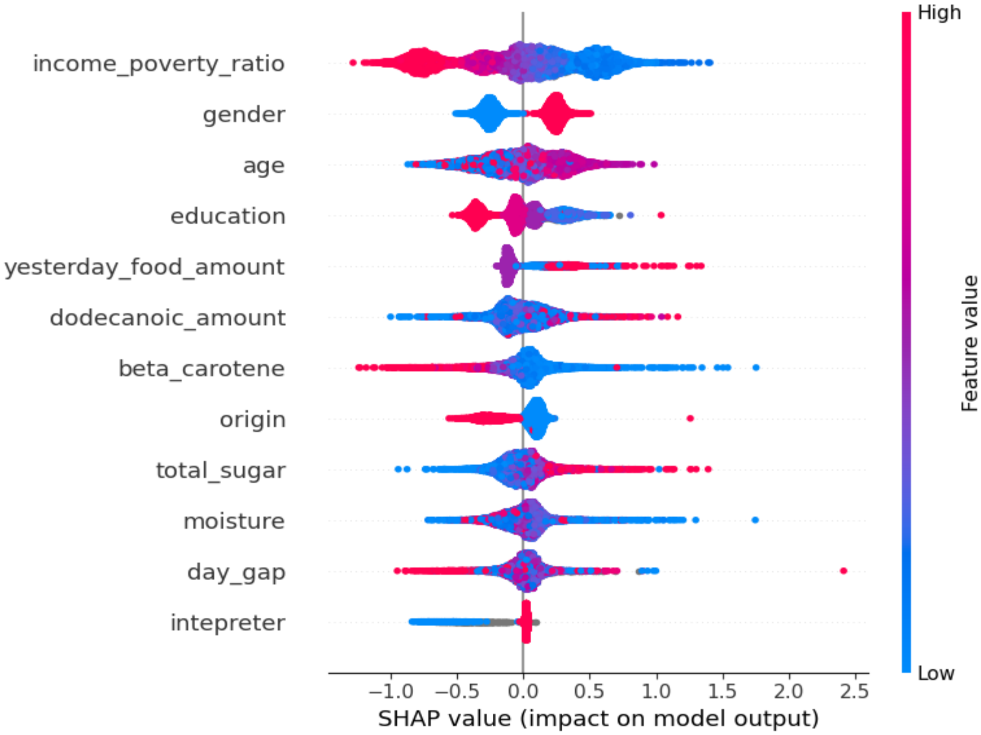

Figure 10 - SHAP value of the final Catboost model

VI. Conclusions and Recommendations

Deducting the results from analysis, it is observed that while the model accuracy is high, the low recall metric underscores a critical shortcoming: the model's limited capacity to accurately identify most positive cases. This limitation poses challenges, especially when the overarching aim is proactive mental health management.

Diving into the model's intricacies, the feature importance and SHAP values elucidate the significant predictors. The income_poverty_ratio stands out as a paramount determinant, implying the profound interplay between economic conditions and mental well-being. Concurrently, the dodecanoic_amount demands further exploration, hinting at underlying complexities. Age, education, dietary components, and gender further enrich the model's landscape, each bringing unique nuances to the predictive paradigm.

In summary, while our model offers valuable insights into the predictors of depression, its current form may benefit from further refinement, especially to enhance its recall. The goal remains unwavering: fostering a tool that aids in the timely and effective identification and intervention for those grappling with or at risk of depression.

From my point of view, for future improvement of the project objectives, there are three potential approaches. Firstly, as the data is highly imbalanced, it is necessary to balance it using balancing techniques such as oversampling, undersampling, or the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes. In my case, I have attempted to use these techniques with subjective parameters; however, the result was not improved. Secondly, the Depression Status might be defined in different ways, as known as the cutpoint of the Depression Score to decide that the participant has depression or not. As stated, I defined the Depression Status with the PHQ-9 approach and there still may be many ways that could potentially be more suitable for defining the label. Lastly, owing to the rapid advancements in technology, the emergence of even more sophisticated machine learning algorithms in the future could discover more patterns in data and provide better advanced performance to reach the project objectives.

VII. References List

CDC. (2015). *NHANES 2015-2016: Mental Health - Depression Screener Data Documentation, Codebook, and Frequencies*. gov.cdc.wwwn. Retrieved September 5, 2023, from https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DPQ_I.htm#Component_Description

Chanthadavong, A. (2021, June 28). *WHO warns against applying AI models using data from rich countries to everyone else*. ZDNET. Retrieved September 5, 2023, from https://www.zdnet.com/article/who-warns-against-applying-ai-models-using-data-from-rich-counties-to-everyone-else/

LaKind, J., Goodman, M., & Naiman, D. (2012, December 5). *Use of NHANES Data to Link Chemical Exposures to Chronic Diseases: A Cautionary Tale*. NCBI. Retrieved September 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3515548/

Muza, C. (2021). *Why you should analyze the distribution of your Data*. Devansh- Machine Learning Made Simple. Retrieved September 5, 2023, from https://machine-learning-made-simple.medium.com/why-you-should-analyze-the-distribution-of-your-data-695fd9f0f1be

NCI. (2022). *Oral Microbiome Analyses Data from NHANES - NCI*. Division of Cancer Epidemiology and Genetics. Retrieved September 5, 2023, from https://dceg.cancer.gov/research/how-we-study/microbiomics/nhanes-oral-samples

NHANES. (2020). *Data Access - Data User Agreement*. CDC. Retrieved September 5, 2023, from https://www.cdc.gov/nchs/data_access/restrictions.htm

NHANES. (2022). *NHANES - NCHS Research Ethics Review Board Approval*. CDC. Retrieved September 5, 2023, from https://www.cdc.gov/nchs/nhanes/irba98.htm

NHANES & CDC. (2021). *NHANES - Participant - Your Privacy*. CDC. Retrieved September 5, 2023, from https://www.cdc.gov/nchs/nhanes/participant/participant-confidentiality.htm

Wade, D. (2007). *Ethics of collecting and using healthcare data*. NCBI. Retrieved September 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1906611/

Williams, J. (2001). *The PHQ-9: validity of a brief depression severity measure*. PubMed. Retrieved September 5, 2023, from https://pubmed.ncbi.nlm.nih.gov/11556941/