A Comparative Analysis and Predictive Model for Reporting Depression in U.S. Adults

Kiran Mai Jaiswal Charpuria, Kruthika Gaddam, Mohith Surya Kiran Kasula, Bala Samantula, Sri Harsha Sudalagunta, April Taylor, Alan Varkey

Indiana University-Purdue University, Indianapolis, USA

[kichar, kgaddam, mkasula, bsamantu, ssudala, taylorad, alvarkey] @iu.edu

Abstract. A comparative analysis and predictive model for the reporting of depression in adults was performed. The objectives of this study were 1) evaluating discrepancies between self-reporting and healthcare provider reporting of depression and 2) to develop a machine learning algorithm to predict specific groups most likely to be classified with depression. Data from two existing publicly available datasets was utilized. Based on a Depression Indicator, 24.5% (1436/5856) of adults reported depression while 15.4% (1344/8701) of adults were reported with depression by providers.

A binary classification model was used on both datasets. Among the 2 machine learning models, the XGBOOST model had the best fit compared with the others (NAMCS: accuracy: 0.850; precision: 0.549; sensitivity recall: 0.51, specificity: 0.918; precision- recall curve or AUC 0.569, F1_score: 0.5292 and NHANES: accuracy: 0.738; precision: 0.449; sensitivity recall: .0.307, specificity: 0.878; precision- recall curve or AUC: 0.463, F1_score: 0.3644.

This study provides preliminary evidence for developing a machine learning program to predict the classification of depression in adults. The top features predictive of depression were persons with no chronic conditions, cardiopulmonary diseases, cancer or blood related diseases, arthritis, or overweight. Further study is recommended based on these findings for the development and testing of a screening algorithm for providers.

Keywords: Depression; Reporting; NHANES database; NAMCS database; Predictive model

1 Project Scope

1.1 Introduction

People with mental health disorders are at an increased risk of social, educational, racial, and physical difficulty[1]. The incidence of depression globally in 2019 was 280 million, and the World Health Organization reported an estimated 28% increase in 2020 due to the Covid pandemic[2]. The American Academy of Pediatrics, American Acad-

emy of Child and Adolescent Psychiatry, and Children's Hospital Association declared a national mental health emergency for children and adolescents in the United States[3].

Despite the rise in depression and the effects of the disorder across all age groups, there remains a personal and medical stigma to a depression diagnosis[4]. Even mental healthcare providers have demonstrated stigma toward patients with mental health disorders[5]. Stigma, whether self-imposed or experienced, particularly affects the disclosure of mental health issues. Garcia et al.[6] demonstrated that standardized screening increased the diagnosis of depression in disparaged groups. Appropriate screening in an emotionally safe healthcare setting is a critical first step to addressing the global issue of depression.

The focus of this project was the screening and diagnosis of depression in adults. Due to access limitations to depression data, minors were excluded. The objective of this comparison was to analyze the reporting behaviors of the adult population and the diagnosis behaviors of healthcare providers regarding depression and then to create a model for predicting depression.

1.2 Aims

- To evaluate any significant differences between self-reported and provider-reported cases of depression in adults in the U.S.
- To identify and correlate the main factors impacting the reporting of depression by self-reporters and providers in adults in the U.S. and identify the groups that report depression the most.
- To develop a multi-variant model that predicts the diagnosis of depression cases focusing primarily on the features of age, gender, income, co-morbidity, medical insurance, education level, and race.

1.3 Purpose

This project investigated differences between self-reporting and healthcare provider reporting of depression, and the findings were used to predict specific groups, for example, a specific age range, who were more likely to experience depression. The results can be used to improve screening standards for healthcare providers to increase the diagnosis of depression in groups predicted to have depression.

1.4 Hypotheses:

Null Hypothesis. There will be no increase in the reporting of depression by adults as compared to the reporting of major depressive disorder by healthcare providers in medical records in the US.

Alternate Hypothesis. There will be an increase in reporting depression by adults compared to major depressive disorder by healthcare providers in medical records in the US.

2 Methodology

This project was a comparative analysis between adults and healthcare providers. The project used two datasets to compare a Depression Indicator from the NHANES dataset with a Depression Indicator from the NAMCS dataset. Once the data was compared, machine learning models were developed to predict the factors impacting the Depression Indicator in 1) adults and 2) healthcare providers. Tools for the project included Python Jupyter notebook, phpMyAdmin, Google Colab, and Microsoft Teams.

2.1 Stages of Project:

1. Data Collection and Extraction
2. Data Cleaning and Storage
3. Data Analysis
    a. Exploratory Data Analysis
    b. Prediction Modeling
4. Data Visualization

2.2 Team members Contributions

Everyone was incredibly supportive on our team, as everyone participated in the project actively. We met every Friday at 11 A.M. For those that could not meet, notes were provided on Canvas to keep track of the project and for the professor and TAs to monitor our progress and provide feedback. Meetings were scheduled weekly to remain in touch and obtain a corporate feel, which was a positive aspect of this.

Table 1.

Team Members							
Tasks	**April Taylor**	**Kruthika Gaddam**	**Bala Samantula**	**Sri Harsha Sudalaguntа**	**Mohith Surya Kiran**	**Kiran Mai Jaiswal**	**Alan Varkey**
Project Management	■	■	■			■	■
Background/ Research	■						
Proposal Development	■	■	■	■	■	■	■
Editing/ Proofreading	■	■	■	■	■	■	■
Data Collection	■						■
Data Analysis		■	■		■	■	
Hypothesis Testing	■	■		■			■
Project Presentation	■	■	■	■	■	■	■
Report Development	■	■	■	■	■	■	■
Data Cleaning	■	■	■		■	■	
Model Development and Testing				■			■
Data Visualization		■	■	■	■	■	■

Canvas was utilized efficiently for discussions to conclude while considering everyone's viewpoints. In addition to the lecture materials and videos, we exchanged YouTube videos in our group as we came from diverse backgrounds and had varying levels of knowledge about Python, ML, GitHub, and Colab. Microsoft Teams was effectively used to upload files and record meetings. We utilized Wrike, a project management platform, to keep track of activities because we were divided into sub-teams and had weekly responsibilities to complete.

2.3 Data Description

Data from two existing publicly available datasets was utilized for this project. Both datasets consisted of secondary data collected by the National Center for Health Statis-

tics (NCHS) and contained data from surveys and physical examinations of persons in the United States (U.S.) from 2017 to 2018[7]. The project used 2018 survey data for both datasets because more recent full datasets were limited due restrictions to in-person contact during the Covid pandemic. The project team also considered the possible compounding effect the Covid pandemic may have on depression rates and decided to assess the pre-pandemic state of depression.

NHANES. The NCHS routinely conducts the National Health and Nutrition Examination Survey (NHANES) to gather information on the health of the general U.S. population. Nutrition, disease, health behaviors and demographic information are among the data collected from participants through a "mobile examination center" [8]. The 2018 source dataset included data from 9,254 persons[9].

NAMCS. The NCHS also routinely conducts the National Ambulatory Medical Care Survey to gather information on the direct patient care received in ambulatory care centers. Examples of available data from this dataset include patient diagnosis codes, insurance coverage, provider types and practice descriptors[10]. The 2018 source dataset included data from 9,953 patient record forms submitted by 496 physicians[11].

2.4 Data Collection and Extraction

Collection. The NHANES dataset is sectioned into multiple subset data files based on topic or survey instrument. The subset datasets are available from the NCHS website as XPORT files (xpt) along with codebooks for interpretation[8]. The subsets which included variables pertaining to this project were chosen. The NAMCS data is available on a NCHS website in a SAS file (sas7bdat) in a zip file. The NCHS provides a microfile data codebook for interpretation[12]. The source files were downloaded and stored in a shared repository (Teams) and the class repository (Canvas). The source files were not stored directly in MySQL due to error messages when attempting to upload the files in their original state.

Table 2.

PROJECT SOURCE FILES	
2018 NHANES	
Demographics	DEMO_J.XPT
Disabilities	DLQ_J.XPT
Depression Screener	DPQ_J.XPT
Medical Conditions	MCQ_J.XPT
Health Insurance	HIQ_J.XPT
2018 NAMCS	
Office-based visits	namcs2018_sas.sas7bdat

The files were uploaded into Jupyter, for processing with Python. The SAS and xpt files were converted to csv files using the Python Pandas method. The 5 NHANES CSV files were merged into one CSV file using a Python pandas merge method with parameters for an outer merge. The SEQN column was used as a primary key in each set to allow for the retention of all rows and columns from all five datasets. Since not all surveys were conducted on all participants, the merging of the datasets led to multiple empty cells in the merged NHANES set. The empty cells were filled with '0' to aid in the data analysis process.

Extraction. Both surveys include a broad amount of health data. This project did not analyze the full datasets. It focused on the data within each dataset that was anticipated to correlate with depression or depression reporting. The datasets were reduced with the pandas method of selecting only the columns designated for this project. Medications were also excluded from both datasets, except for the Boolean variable of "medication for depression." Medications are varied in their indication for use, and the datasets did not provide information on initiation or discontinuation dates for correlation with depression onset or improvements. Due to this complexity, the team chose not to include this variable in the project, despite the potential to indicate or predict depression.

Removing irrelevant data. The data was further reduced to only data relevant to the project aims. After merging the NHANES subsets, the datasets were cleaned of duplicates.
AGE. Age was limited to adults only in both sets. The NCHS restricts the release of depression scores for minors in the NHANES survey, therefore participate data for those under age 18 was removed for all variables. Age was limited in the NAMCS dataset with the selection of only the column, "AGE" which limited the participant's data to those aged over 17.

Handling Missing Data. *NHANES.* Per the NCHS recommendation[13], data was evaluated for missing values and as prespecified any variables with more than 10% missing values would be imputed. The following values in the set indicated no data was available:

- missing = blank
- "don't know" = 9/99
- refused = 7/77

Table 3.

For all variables in the reduced dataset, the above were considered in the missing data calculations. Not all survey questions were asked or appropriate for all participants. For those variables, blanks were converted from NaN to "0".

NAMCS Missing Values(%)			NHANES Missing Values(%)			
Code	**-9**		**Code**	**9/99**	**7/77**	**Blank**
RACER	0%		INDHHIN2	0.1%	2.0%	0.0%
RACEUN	28.6%		HIQ031A	0.7%	0.1%	0.0%
PAYTYPER	2.5%		HIQ031B	0.0%	0.0%	0.0%
DIAG1	0.4%		HIQ031D	0.0%	0.0%	0.0%
DIAG2	34.2%		HIQ031E	0.0%	0.0%	0.0%
DIAG3	57.9%		HIQ031H	0.0%	0.0%	0.0%
DIAG4	73.1%		HIQ031I	0.0%	0.0%	0.0%
DIAG5	84.3%		HIQ031AA	0.0%	0.0%	0.0%
OWNSR	4.5%		DMDEDUC2	0.0%	0.2%	0.0%
AGE	0.0%		DMDEDUC3	0.0%	0.0%	0.0%
SEX	0.0%		DLQ140	0.0%	0.0%	1.6%
DEPRN	0.0%		DLQ150	0.1%	0.0%	1.6%
NOCHRON[1]	67.0%		DLQ170	0.2%	0.0%	48.0%
DEPRESS	0.0%		DPQ010	0.1%	0.7%	7.5%
MENTAL	0.0%		DPQ020	0.1%	0.0%	7.5%
PSYCHOTH	0.0%		DPQ030	0.1%	0.5%	7.5%
NOPROVID	0.0%		DPQ040	0.1%	0.1%	7.5%
PHYS	0.0%		DPQ050	0.1%	0.0%	7.5%
PHYSASST	0.0%		DPQ060	0.1%	0.1%	7.5%
NPNWM	0.0%		DPQ070	0.1%	0.0%	7.5%
RNLPN	0.0%		DPQ080	0.1%	0.0%	7.5%
MHP	0.0%		DPQ090	0.1%	0.0%	7.5%
OTHPROV	0.0%		RIDAGEYR	0.0%	0.0%	0.0%
PROVONE	0.0%		RIAGENDR	0.0%	0.0%	0.0%
SPECCAT	0.0%		RIDRETH3	0.0%	0.0%	0.0%
	[1] Blank		>10%			

PHQ-9 Questions. The blank data in the PHQ questions was reviewed in detail to ensure the blank values that were converted to "0" for the depression screening did not skew the PHQ total score to lower values or the overall rate of depression in NHANES. Each question for the PHQ had a 7.5% missing rate. This consisted of about 743 participants for which all questions were blank for the PHQ-9. This indicates the participants did not consent to the depression survey. The team chose not to drop these participants to preserve the other data obtained for the other variables (including other Depression Indicators) also considering that the missing data rate was less than the prespecified 10%.

As further discussed in Descriptive Statistics., the NHANES dataset Depression Indicator proportion was higher than the NAMCS total Depression Indicator proportion. The team did not feel the missing surveys responses significantly affected the outcome of the findings.

DLQ-170. How depressed did you feel? This question was an addendum question added to the PHQ-9 and the only other variable with more than 10% missing within NHANES. The team chose not to impute this variable considering that a participant may feel uncomfortable or uncertain answering this question. Also, a positive response was cap-

tured in the Depression Indicator. Imputing may falsely increase the dependent variable.

NAMCS. . For most of the variables in the NAMCS, the data was imputed by NAMCS, therefore no adjustments were required. Diagnosis (including no chronic illness), payer type and practice type were not imputed.

Race. Initially, the variable for race, "RACEUN" was chosen, however, it was discovered that 29% of that variable was missing and was not imputed. On further investigation, the "RACER" variable was an acceptable alternative. The "RACER" variable contained less categories within racial groups (black, white, other) which provides less insight into the depression of multiple races. However, the "RACER" variable is imputed by NCHS, and the team agreed having a complete dataset for this variable was more important for this project considering the risk of limited findings from missing values.

Diagnoses. Diagnosis codes, "DIAG" 1-5, were not imputed by NCHS. Due to the nature of U.S. office visits scheduling, most patients are seen for short appointments where only a primary diagnosis (DIAG1) is treated. Rarely are up to 5 diagnoses treated in one office visit (DIAG 1-5). Imputing a diagnosis would introduce inaccurate findings considering the 339 diagnosis categories and numerous subcategories in each diagnosis category. Missing percentages were calculated for these variables, but empty cells were not imputed. Positive findings for diagnoses were used for each participant in the data analyses and predictive models.

No Chronic Illness. A significant amount of missing data (67%), for "no chronic illness" was found, however this category would be difficult to impute accurately considering the wide range of chronic illnesses that could be imputed. As with the depression survey, the team chose not to drop these participants to retain the other data obtained from these surveys.

2.5 Data Cleaning and Storage

Feature Engineering. After cleaning, the datasets were prepared for data analysis by conducting feature engineering. Most of the variables were categorical and were transformed into Boolean (0/1) features. Several variables were combined or transformed by normalization or standardized across datasets for consistency and simplicity. The variables combined and/or renamed include insurance, education, age, gender, and race.

Education. The NHANES source dataset included two education variables; under age 20 and over age 20. Considering that our sample population was all adults over 17, the team created an EDUC feature that merged the two variables of education and consolidated the multiple primary education categories into one category. This allowed for a simplified 5 category education feature.

Insurance. Health insurance was a common variable between the NHANES and NAMCS datasets. However, the datasets differed in the categorization of the types of

insurance. The insurance columns for each dataset were consolidated to 4 common U.S. insurance types: private, state-funded, Medicare, and no insurance.

Race. Race was also a common variable between the datasets. However, NAMCS categorized the imputed RACER variable into three categories while NHANES used 6. For standardization, NHANES categories were consolidated into black, white, and other and both datasets were renamed to have consistent features names.

Comorbid Conditions. Both datasets included medical diagnoses information. NHANES survey captured Boolean answers for chronic medical conditions. The team chose chronic conditions that were anticipated to correlate with depression reporting. The NAMCS survey captured all diagnosis codes as described in the prior section. However, each code was grouped per participant and per visit. The team separated the diagnosis codes into 16 medical condition features using Python pandas as a yes/no Boolean feature for each NAMCS participant. The diagnoses codes for depression were excluded from these features and were used in the creation of the Depression Indicator (see below).

Normalizing depression data. The depression data was normalized to a single depression classification by creating a the " Depression Indicator" (DI) feature for each dataset. For each participant, any positive dependent variable (depression) was considered a "yes" for the Depression Indicator. For example, if a participant received a PHQ total of 10 or more and reported having a chronic condition of depression, that participant was considered as a "yes" for reporting depression. If a provider reported a patient with a depression diagnosis code and depression medication treatment that patient was considered as a "yes" for reporting depression. Any indication of depression was classified on a binary (yes/no) scale for standardization across the datasets. The variables from each set to compile DI included:

- NHANES
  - PHQSCR: PHQ Total>10-- (indicative of depression)
  - DLQ140: Feelings of depression-- >=Monthly,
  - DLQ150: Medications for depression-- Yes
  - DLQ170: Depression Level-- A lot, A little or in between a lot and a little
- NAMCS
  - DEPRN-- depression chronic disease
  - DEPRESS-- depression screen
  - MENTAL-- mental health counseling referral
  - PSYCHOTH—psychotherapist referral

Once feature engineering was complete, NHANES contained 26 features and NAMCS contained 41 features used for hypothesis testing.

3 Data Analysis

To view and comprehend our data, Python was used. Many built-in Python libraries were utilized, including pandas, NumPy, Seaborn, Matplotlib, and Scipy. Matplotlib is a Python library for creating interactive visualizations such as bar charts and pie charts.

Seaborn is a visualization library based on Matplotlib used to generate heatmaps. Scipy is a Python package used for scientific computing and hypothesis testing. To help comprehend the data, the visualizations were developed below:

1. A stacked bar visualized associations between the Depression Indicator and:
    a. Age
    b. Race
    c. Gender
    d. Education
    e. Income
    f. Provider types
    g. Co-morbid conditions
2. Pie charts visualized the proportion of DI in NHANES and NAMCS
3. Histograms of Age were plotted to determine each distribution
4. Correlation Heatmaps were created to view and illustrate the correlations between all variables in each dataset
5. Horizontal bar charts displayed the model results in order of feature importance
6. Charts displayed results of confusion matrices
7. PR Curve plots displayed the precision and sensitivity of each model performance
8. ROC plots displayed a comparison of the true positives compared to true negatives and the area under the curve for each model to measure performance

3.1 Descriptive Statistics

Descriptive statistics allow data to be presented more easily, making interpretation meaningful. Since most variables were categorical, only the mean and median were calculated for the continuous variables of age in the NHANES and NAMCS datasets and "PHQ Score" in the NHANES dataset. The mode, standard deviation, variance, and interquartile range were determined for all categorical variables to understand the distribution and dispersion. Skewness and kurtosis were calculated to identify outliers.

wo features were determined to cause large kurtosis. The data was examined and was not found to have outliers. The feature of no insurance only had one participant with a “yes.” This was not unexpected due to the large numbers of participants with other insurances and the changes in U.S. policy recently around insurance access for the uninsured. “No Provider” was the other feature with a large kurtosis. This feature was not a valuable feature as it was only for patients who visited the clinic without seeing a provider. This was not expected to affect the Depression Indicator and was not found to be a significant feature. Normalization of the data was not required for the categorical data therefore was not performed. The results of the descriptive statistics were compiled into a tabular description for easier comprehension (See Appendix).

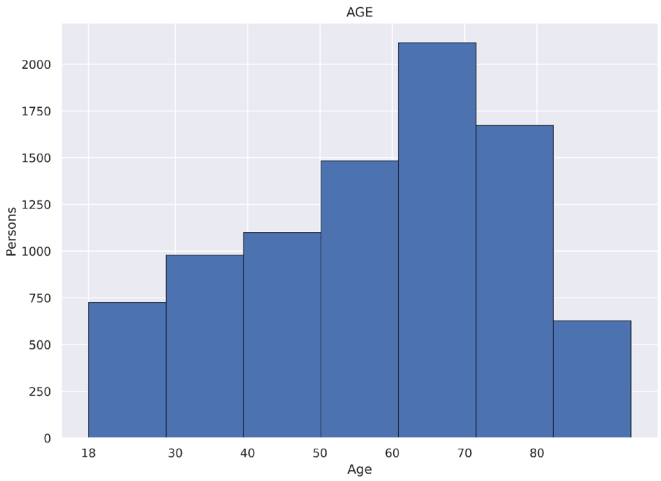

Fig. 1. The NAMCS sample skewed to the older age range with the largest group in the 60-70 age range.

.

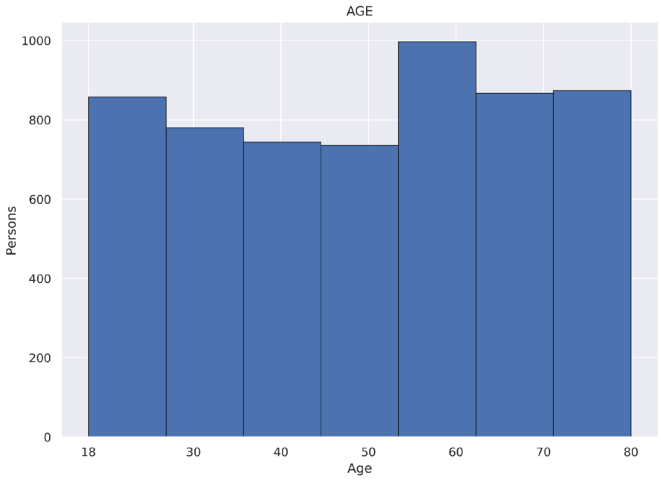

Fig. 2. Age Distribution for NHANES. The NHANES sample was evenly distributed across the age groups with a slightly larger distribution in the 60-70 age range.

Exploratory Data Analysis. Among the two data sets, the common variables were age, gender, comorbid conditions, race, and insurance type. These shared variables were compared across the datasets. Education, income, and provider types were not common variables. All were analyzed for an association with the Depression Indicator.

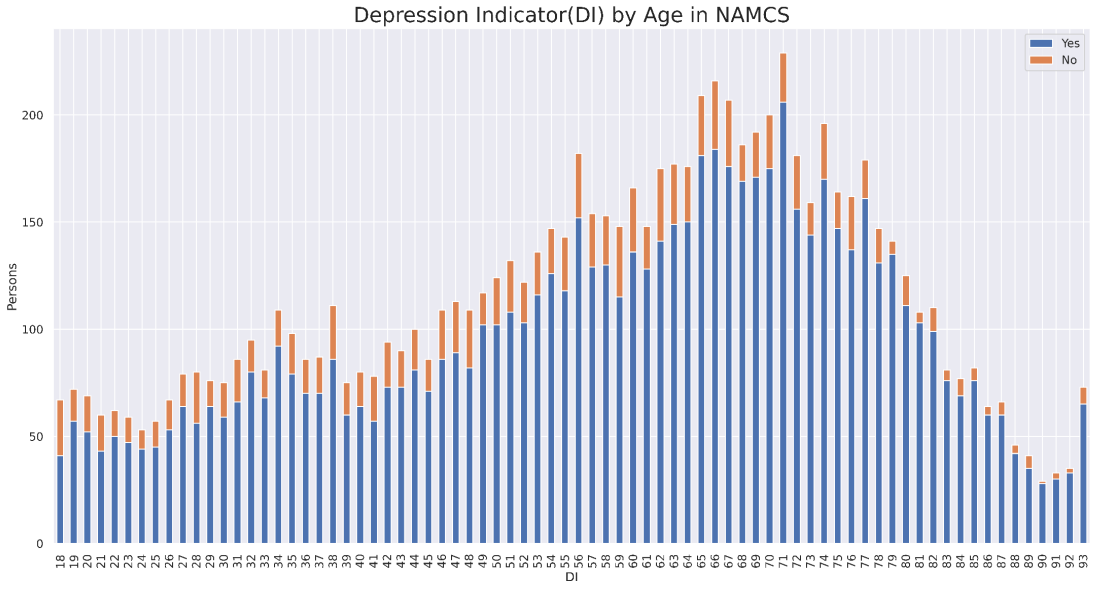

Fig. 3. DI by Age Groups in NAMCS. Higher distribution of depression in the older population and 18-year-olds

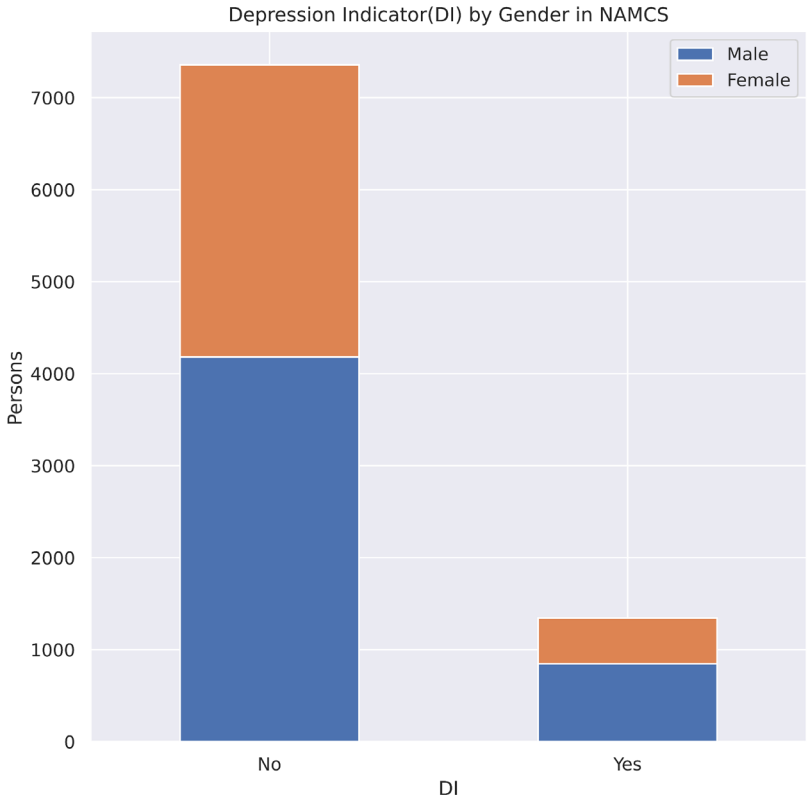

Fig. 4 DI by Gender in NAMCS. Slightly greater DI in males than females.

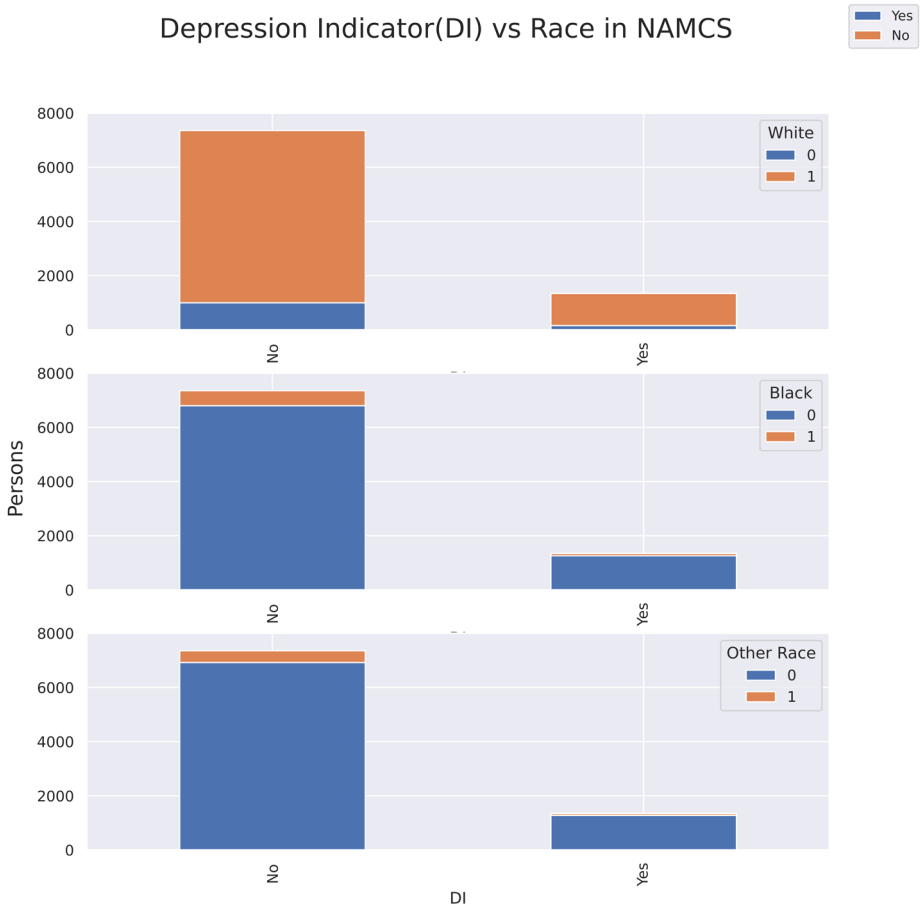

Fig. 5. DI by Race in NAMCS.

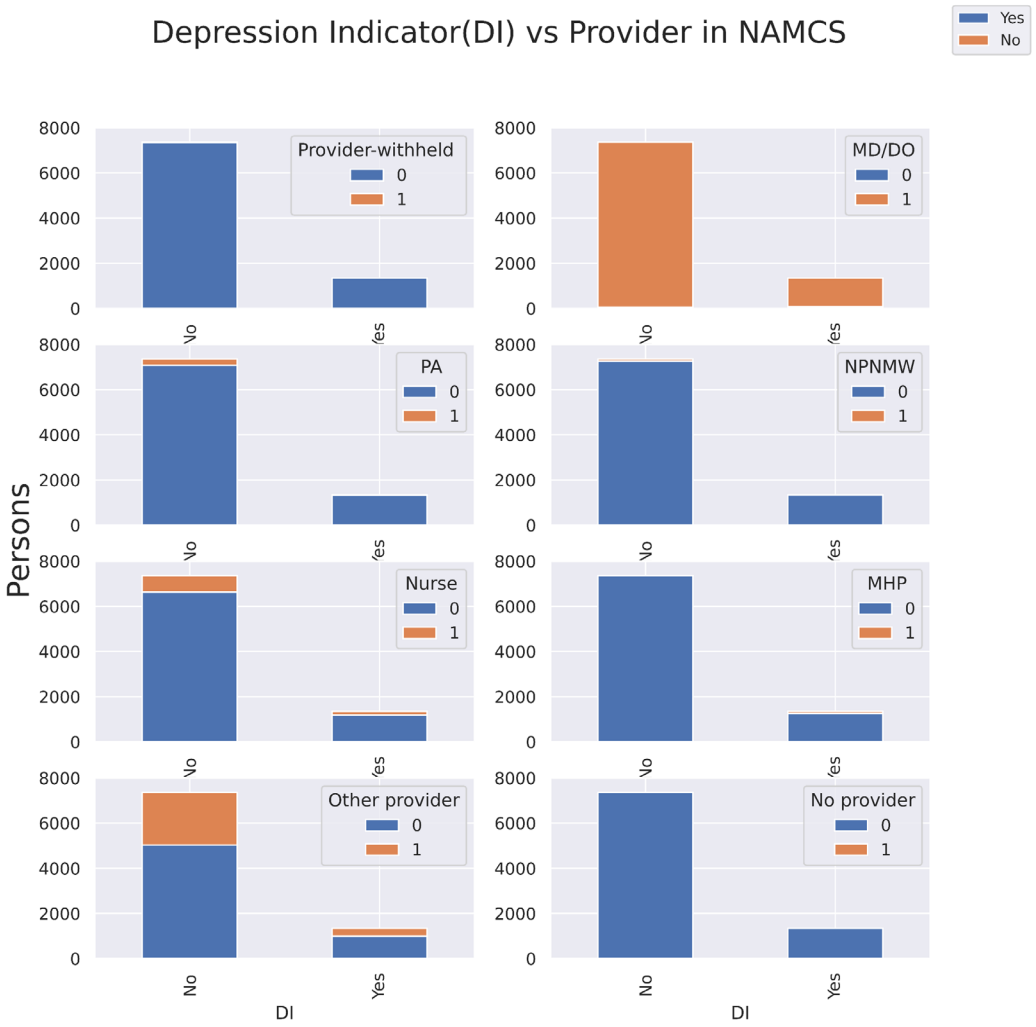

Fig. 6. DI by Provider in NAMCS. Physician type most likely to diagnose depression

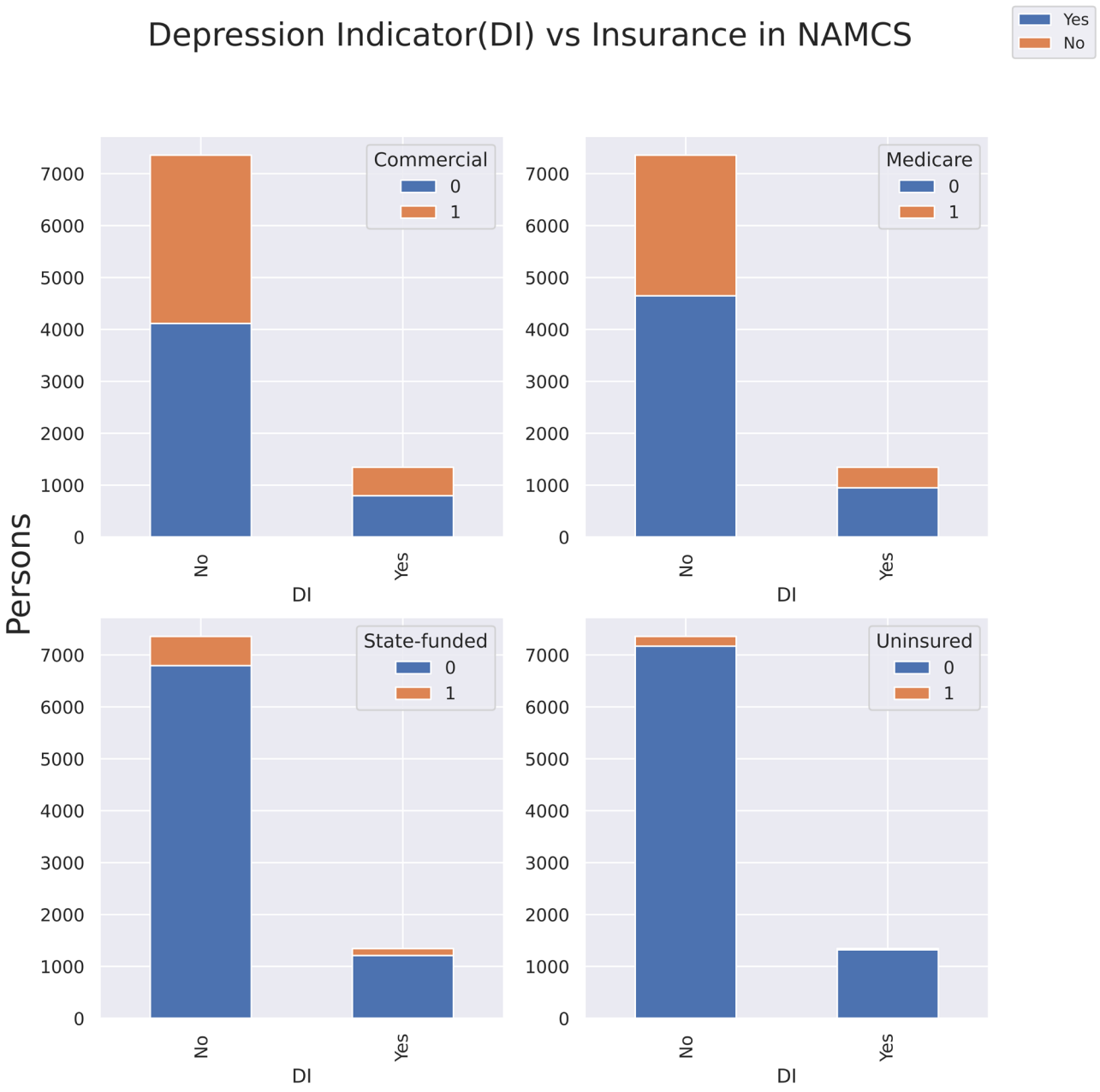

Fig. 7. DI by Insurance Type

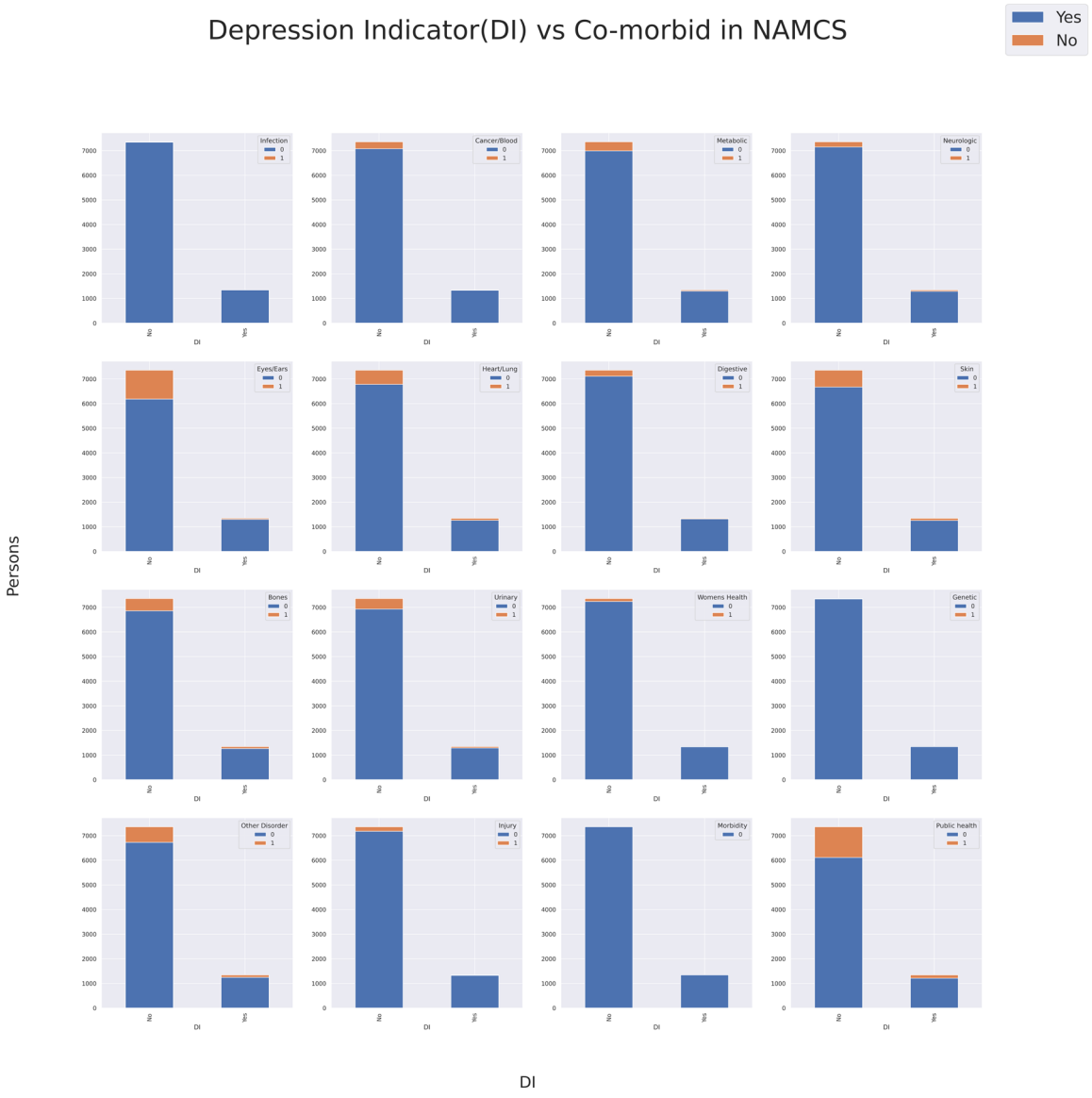

Fig. 8. DI by Comorbid Conditions in NAMCS. Association in those with cardiopulmonary, dermatologic, and public health conditions.

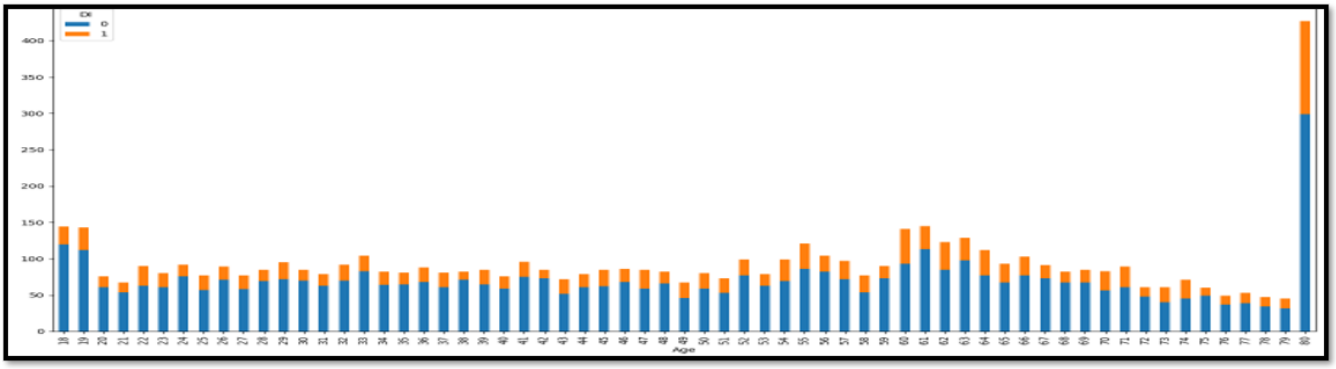

Fig. 9. DI by Age in NHANES. Higher distribution for older and 18–19-year-olds.

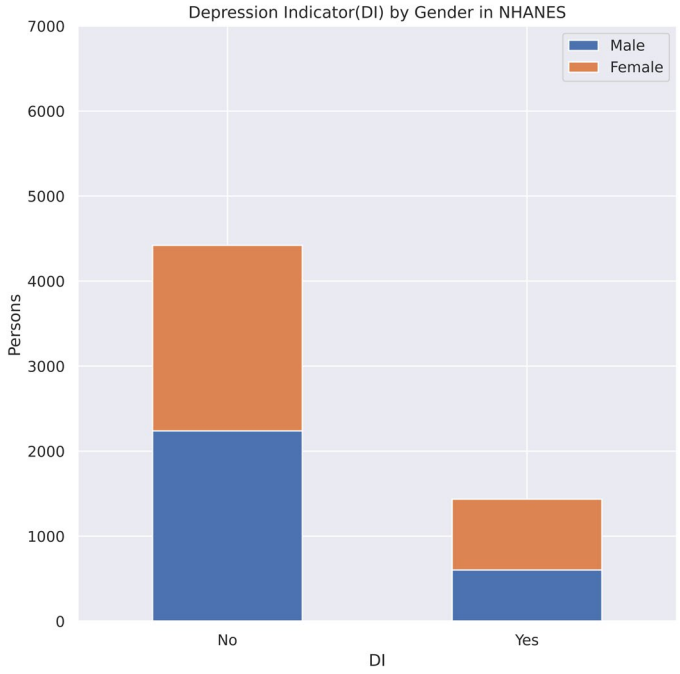

Fig. 10. DI by Gender in NHANES. Slightly more females than males with depression.

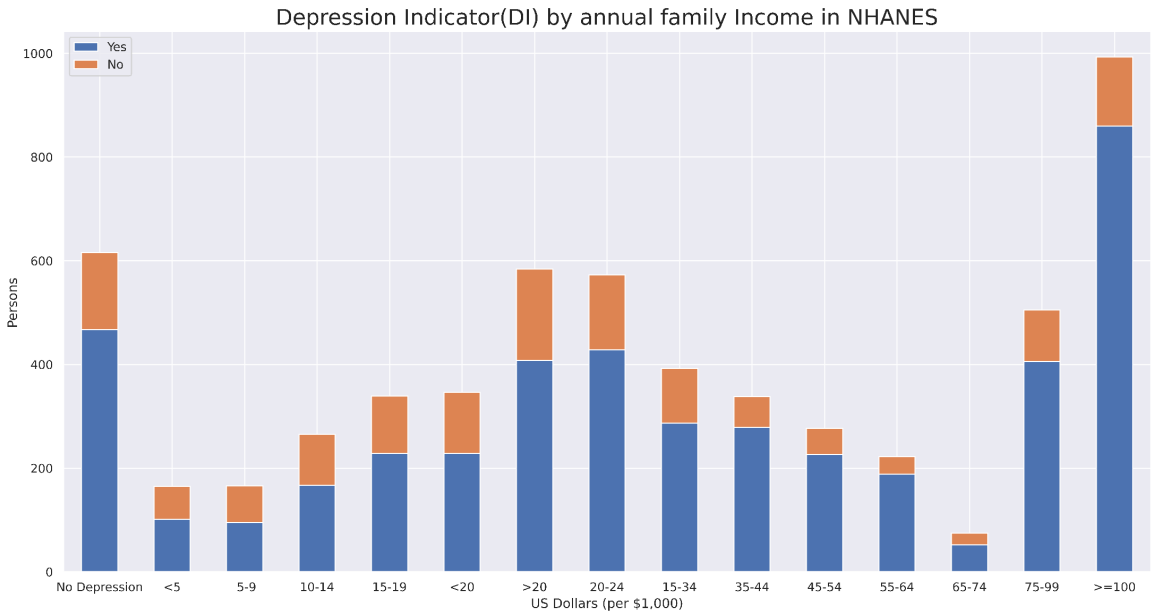

Fig. 11. Largest group with DI is middle income, largest group in sample is >=100,000

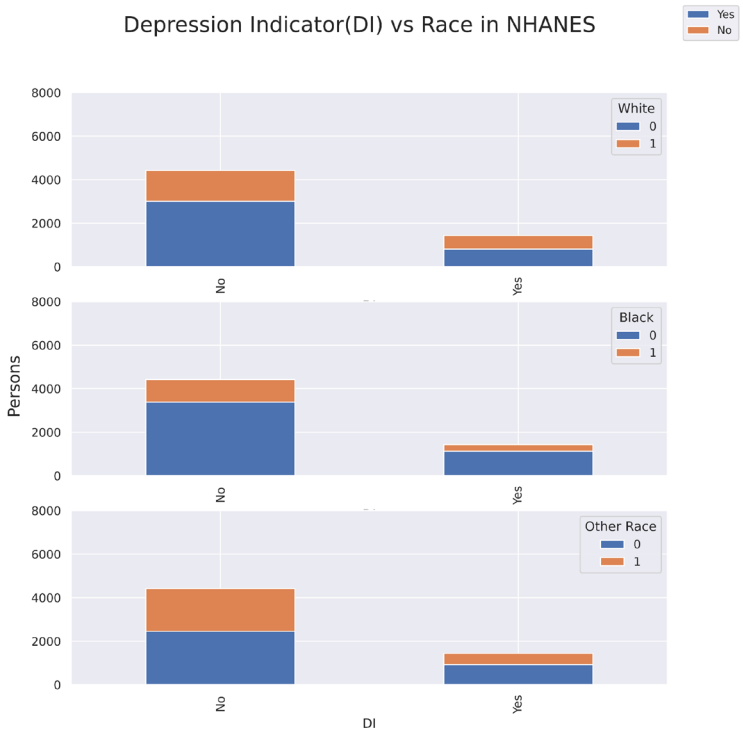

Fig. 12. DI by Race in NHANES

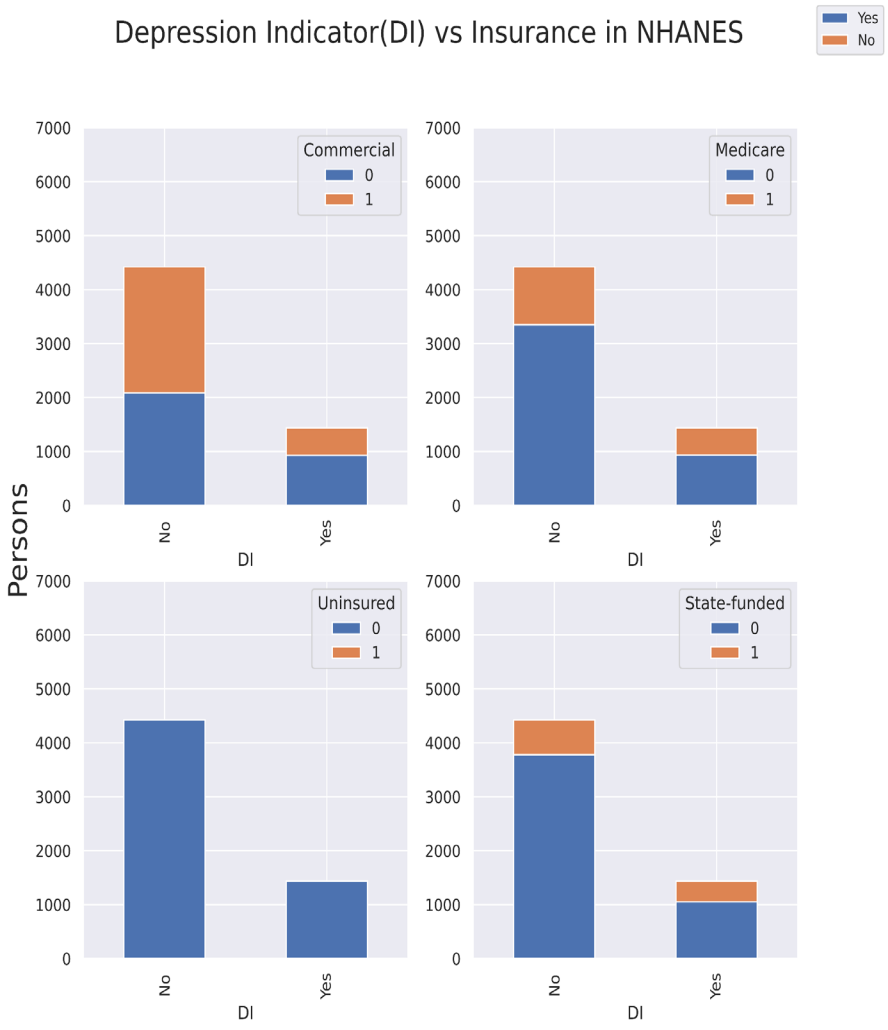

Fig. 13. DI by Insurance in NHANES. Commercial and Medicare have the largest groups with DI and largest distribution of persons in the sample.

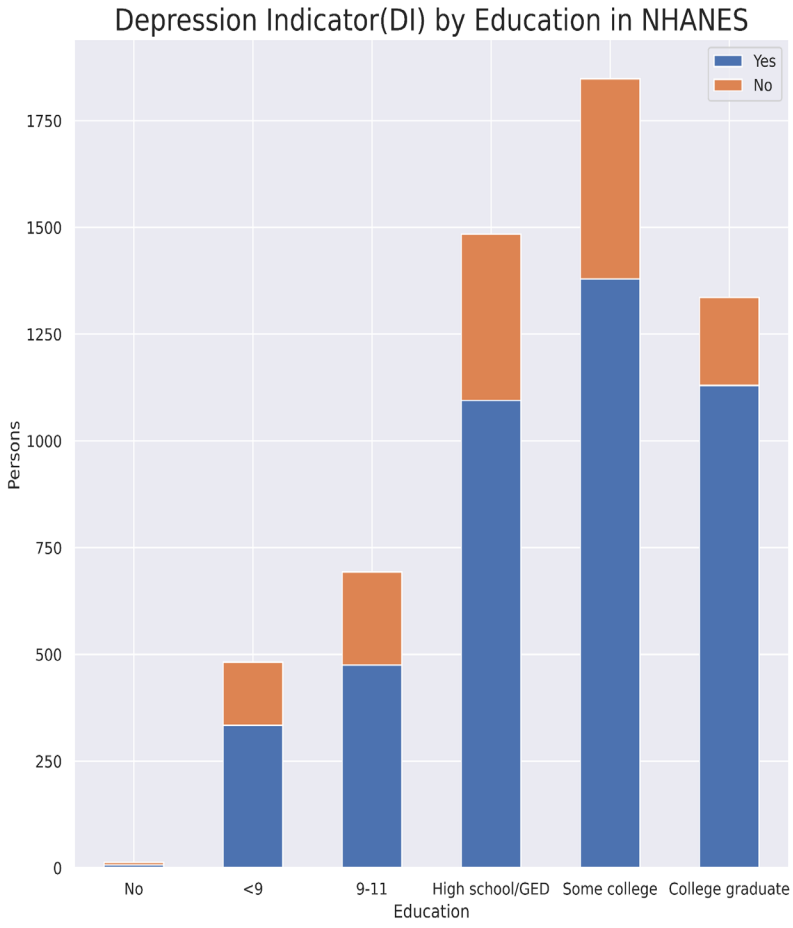

Fig. 14. Education in NHANES.

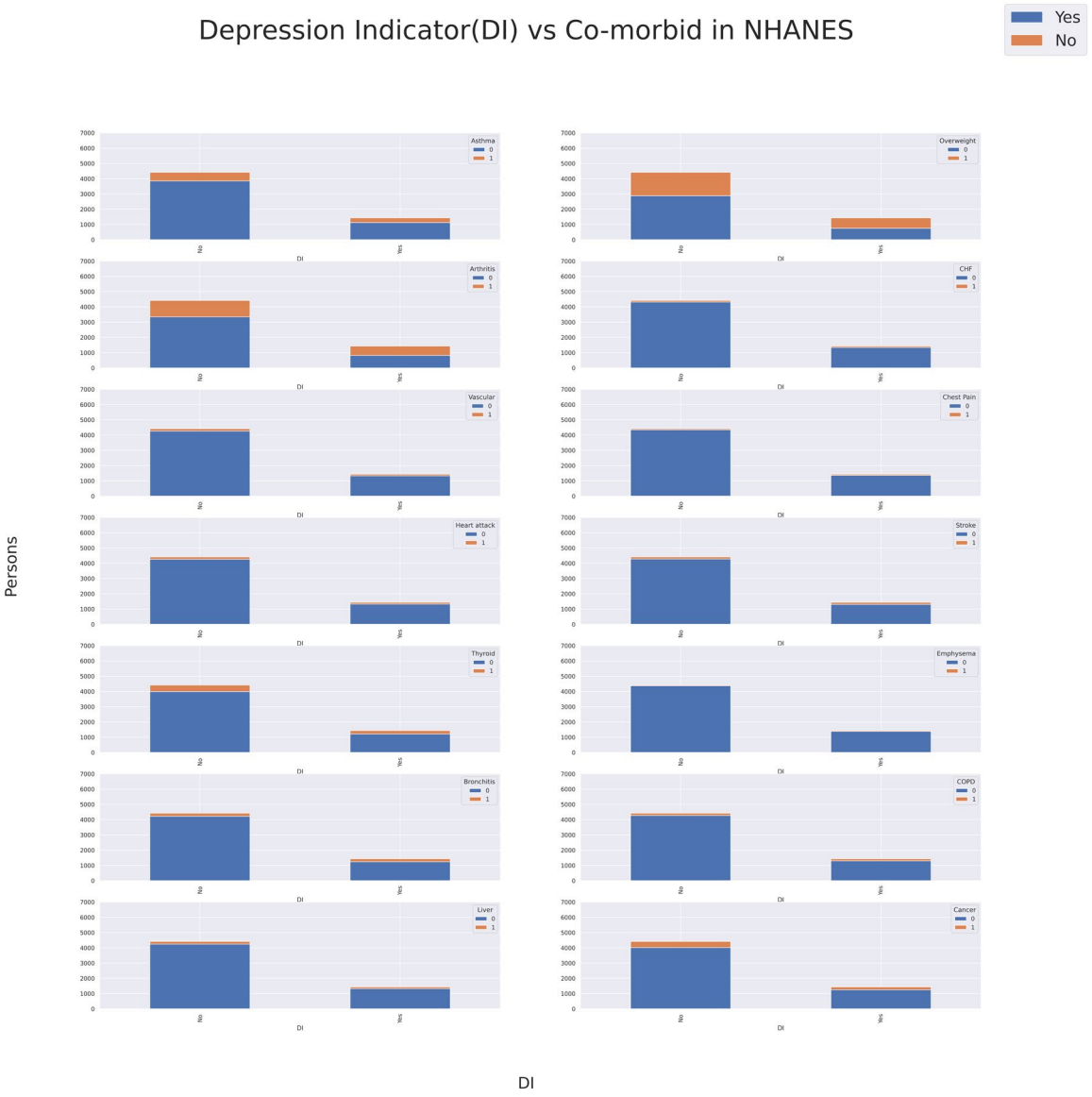

Fig. 15. Comorbid Conditions in NHANES. DI associated with overweight, arthritis and asthma.

3.2 Hypothesis testing

The Chi-square test was used to assess any significant associations between the Depression Indicator (DI) and each independent feature for each dataset to determine the significant features as inputs for the predicative models. For the null hypothesis, a significance of $p < 0.5$ was used as an indication to reject the null and a significant association between the groups. The chi-square contingency test in Python was used for the chi-square analysis. The proportion of the DI between NHANES and NAMCS was also used as a measure of hypothesis testing. The percentage of positive DI per dataset was compared. If the NHANES proportion of DI was higher than the NAMCS proportion, the null hypothesis would be rejected.

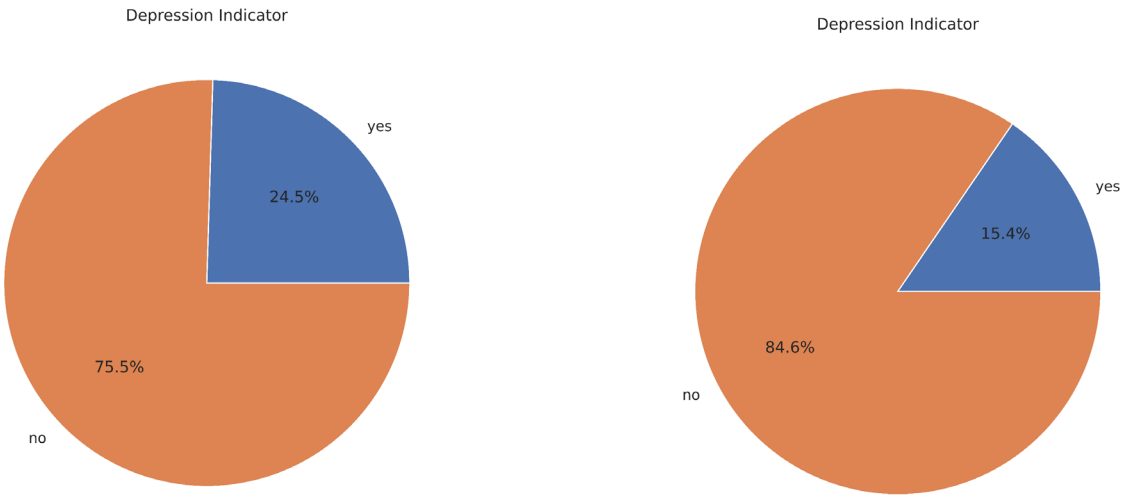

Fig. 16. Proportion of DI in NHANES and NAMCS

The results of the chi-square results determined that 23 and 29 of the features were significant features in NHANES and NAMCS, respectively.

Table 4.

NHANES VARIABLES CHI SQUARE WITH DEPRESSION INDICATOR					
VARIABLE	CODE	P- VALUE	CRITICAL	STATIC	REJECT NULL
Age	AGE	0.03117	81.381	84.336	YES
Income	INDHHIN2	0.00000	23.685	210.484	YES
Asthma	MCQ010	0.00000	3.841	65.169	YES
Overweight	MCQ080	0.00000	3.841	79.925	YES
Arthritis	MCQ160A	0.00000	3.841	182.783	YES
CHF	MCQ160B	0.00000	3.841	56.897	YES
Vascular	MCQ160C	0.00000	3.841	35.056	YES
Chestpain	MCQ160D	0.00000	3.841	55.266	YES
HeartAttack	MCQ160E	0.00000	3.841	37.52	YES
Stroke	MCQ160F	0.00000	3.841	83.849	YES
Thyroid	MCQ160M	0.00000	3.841	37.901	YES
Emphysema	MCQ160G	0.00000	3.841	42.184	YES
Bronchitis	MCQ160K	0.00000	3.841	136.981	YES
COPD	MCQ160O	0.00000	3.841	88.838	YES
Cancer	MCQ220	0.00003	3.841	17.432	YES
Commercial	INSPRVT	0.00000	3.841	130.371	YES
Medicare	INSMCARE	0.00000	3.841	62.943	YES
Uninsured	INSNOCH	1.00000	3.841	0	NO
Gender	SEXML	0.00000	3.841	32.584	YES
White	RACEWH	0.00000	3.841	57.87	YES
Black	RACEBL	0.09907	3.841	2.72	NO
Other Race	RACEOT	0.00000	3.841	34.741	YES
State-Funded	INSSTATE	0.00000	3.841	109.903	YES
Education	EDUC	0.00000	9.488	87.52	YES
Liver	MCQ700	0.00000	3.841	36.454	YES

Table 5.

NAMCS VARIABLES CHI SQUARE WITH DEPRESSION INDICATOR					
VARIABLE	CODE	P- VALUE	CRITICAL	STATIC	REJECT NULL
Age	AGE	0.00000	96.217	198.706	YES
Gender	SEXML	0.00004	3.841	16.681	YES
White	RACEWH	0.04018	3.841	4.21	YES
Black	RACEBL	0.11329	3.841	2.508	NO
Other	RACEOT	0.25653	3.841	1.287	NO
Provider-withheld	NOPROVID	0.41137	3.841	0.675	NO
MD/DO	PHYS	0.00000	3.841	184.963	YES
PA	PHYSASST	0.00001	3.841	20.164	YES
NPNMW	NPNMW	0.33057	3.841	0.947	NO
Nurse	RNLPN	0.09016	3.841	2.871	NO
MHP	MHP	0.00000	3.841	442.487	YES
Other Provider	OTHPROV	0.00005	3.841	16.417	YES
No provider	PROVNONE	0.22218	3.841	1.49	NO
Medicare	INSMCARE	0.00000	3.841	27.031	YES
State-funded	INSSTATE	0.00475	3.841	7.972	YES
Uninsured	INSNOCH	0.11259	3.841	2.517	NO
Commercial	INSPRVT	0.02384	3.841	5.106	YES
Primary Care	SPPRC	0.00085	3.841	11.126	YES
Surgery Care	SPSUR	0.00000	3.841	365.775	YES
Speciality Care	SPMEC	0.00000	3.841	298.753	YES
Owner-withheld	OWNUNKN	0.00448	3.841	8.08	YES
Phy.Owner	OWNPHYS	0.00000	3.841	35.709	YES
Hospital Owner	OWNHOSP	0.00000	3.841	21.41	YES
Insurance Owner	OWNINSR	0.00000	3.841	24.63	YES
No illness	NOCHRON	0.00000	3.841	373.763	YES
Infection	DGRP1	0.50118	3.841	0.452	NO
Cancer/Blood	DGRP2	0.00000	3.841	27.45	YES
Metabolic	DGRP3	0.01191	3.841	6.325	YES
Neurologic	DGRP4	0.02569	3.841	4.977	YES
Eyes/Ears	DGRP5	0.00000	3.841	148.936	YES
Heart/Lung	DGRP6	0.03875	3.841	4.272	YES
Digestive	DGRP7	0.01181	3.841	6.34	YES
Skin	DGRP8	0.00191	3.841	9.631	YES
Bones	DGRP9	0.27198	3.841	1.207	NO
Urinary	DGRP10	0.01331	3.841	6.127	YES
Women's Health	DGRP11	0.00273	3.841	8.983	YES
Genetic	DGRP12	1.00000	3.841	0	NO
Other Disorder	DGRP13	0.16058	3.841	1.969	NO
Injury	DGRP14	0.02574	3.841	4.973	YES
Public Health	DGRP16	0.00000	3.841	40.53	YES

3.3 Feature Correlations

Heatmaps were created to assess relationships between each feature in each dataset. In both datasets, Age and Medicare were strongly associated with each other. This is understandable considering as persons age, the number of citizens that receive Medicare increases in the U.S. The team chose to remove Medicare from the predictive modeling. It was decided that Age represented a broad age range and keeping it would maintain information regarding the younger age groups also.

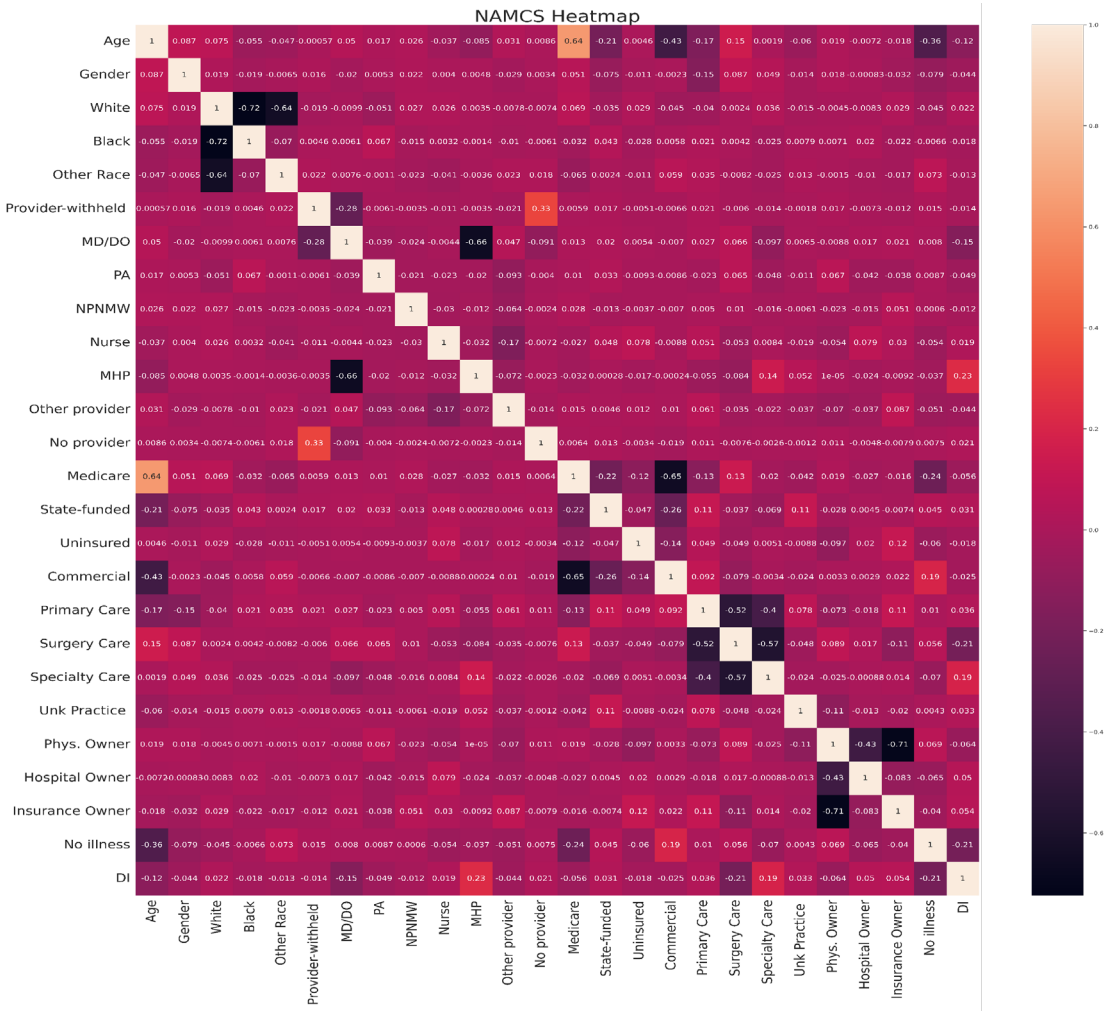

Fig. 17. NAMCS Correlation Heatmap

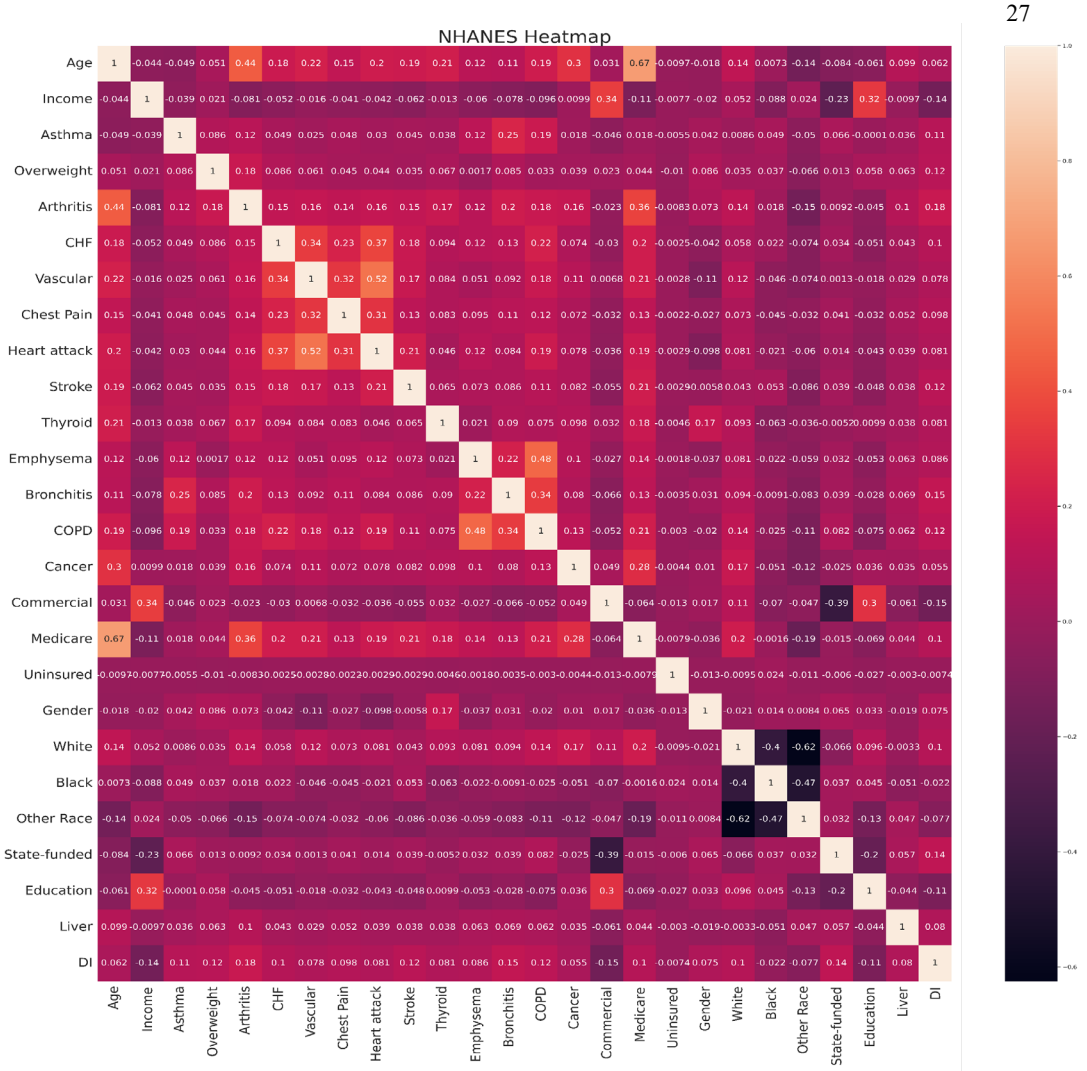

Fig. 18. NHANES Correlation Heatmap

3.4 Machine Learning and Model Testing.

1.1.1 Binary Classification Models.

With the significant variables from both the datasets, two models were used for the prediction of depression among the two datasets. The models that employed were:

- XGBOOST Classifier
- Random Forest Classifier

Each machine learning algorithm was performed on each dataset. Functions used with each model for comparison and performance analysis between the models included:

- Cross-Validation
- Determining Model Metrics
- Feature Importance
- Confusion Matrix
- PR-Curve
- ROC Curve

3.5 XGBOOST Classifier

NAMCS. As 70% of the data represented ‘No” for the Depression Indicator and 30% represented “Yes”, the Synthetic Minority Oversampling Technique (SMOTE) function was performed to balance the data. Testing and training of the NAMCS data set was then performed. The cross-validation score of the NAMCS dataset after SMOTE with XGBOOST classifier using F1_macro was 0.89 with a standard deviation of 0.1. There was a significant difference for the F1 score before and after using the SMOTE.

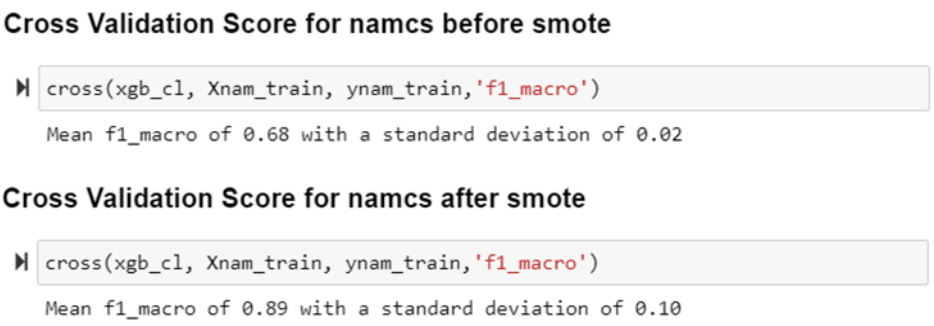

Fig. 19. Before and After Smote NAMCS

The XGBOOST performance scores for NAMCS dataset after SMOTE were:

- Accuracy: 85%
- Precision: 54.9%
- Sensitivity recall: 51%
- Specificity: 91.8%
- F1_score: 0.5292

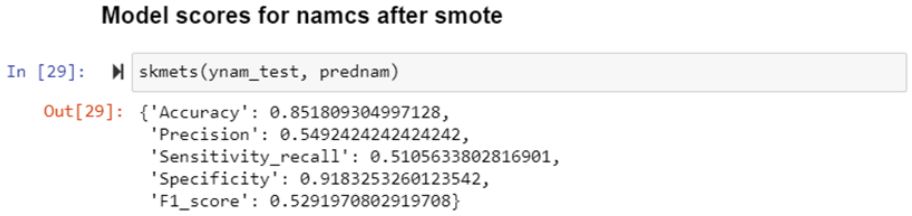

Fig. 20. NAMCS XGBOOST Performance

After using the function SMOTE, the top three features that contributed to the reporting of depression were:

- No chronic conditions
- heart or lung diseases
- cancer or blood related diseases

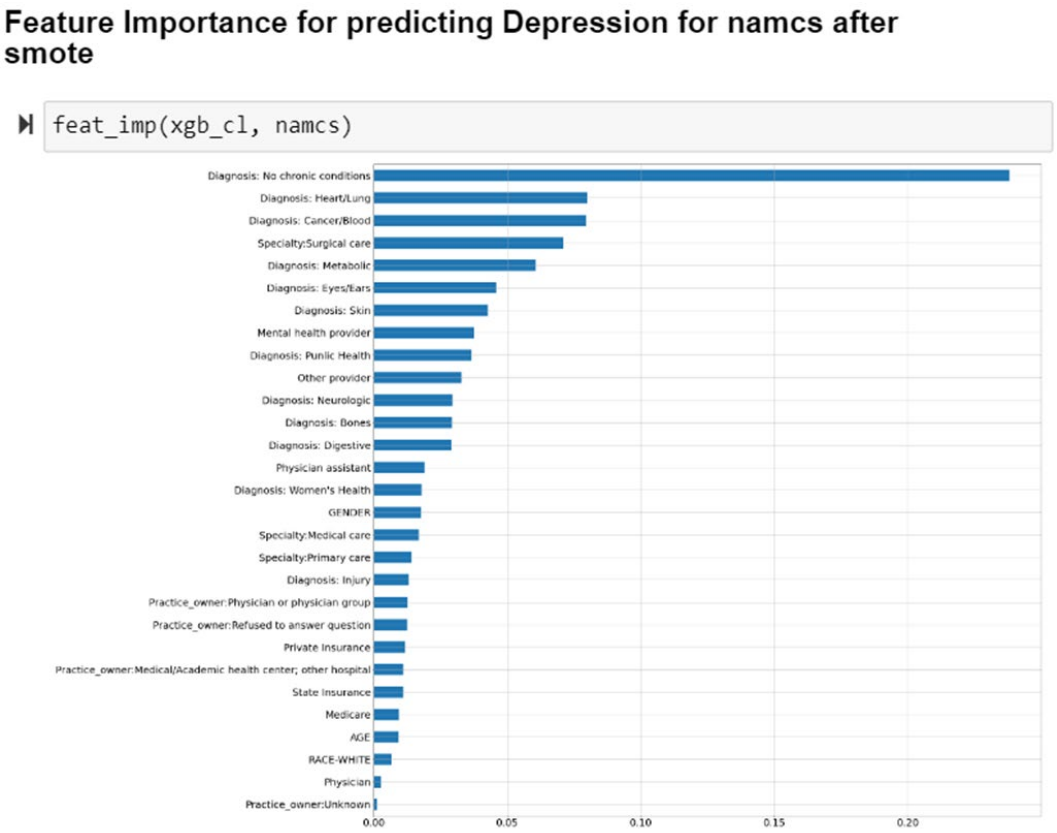

Fig. 21. Fig. 18. NAMCS Feature Importance

Confusion Matrix of XGBOOST classifier. The confusion matrix then provided the model predictions. Of the positive cases from the dataset, the model predicted 51% as True positives and 49% as False Negatives. Of the negative cases from the NAMCS dataset, the model predicted 91.8% as True Negative and 8.2% as False Positives. The model has high accuracy for the prediction of patients with no depression.

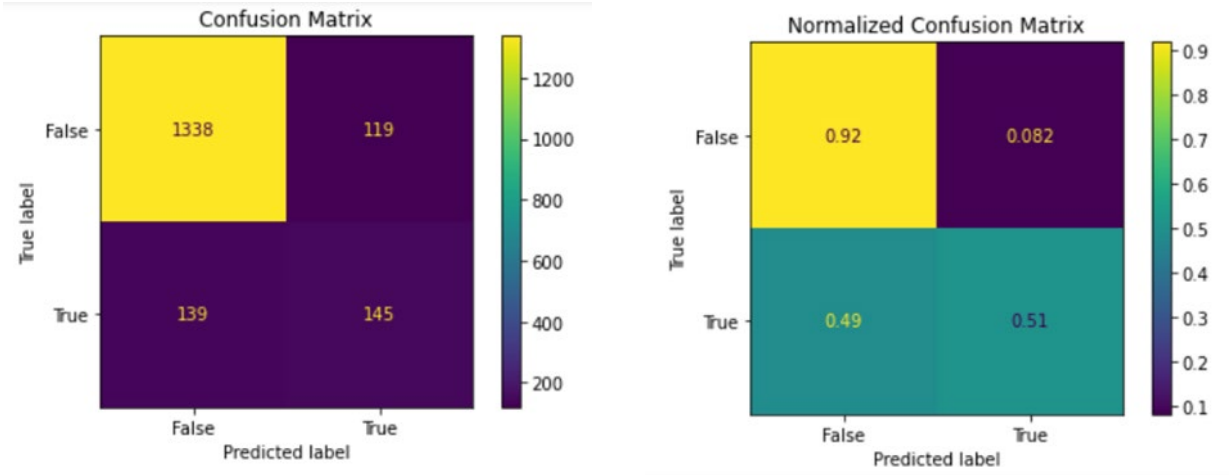

Fig. 22. Confusion Matrix of NAMCS

PR curve. The precision- recall curve for the NAMCS dataset after applying SMOTE, AUC was 0.5698

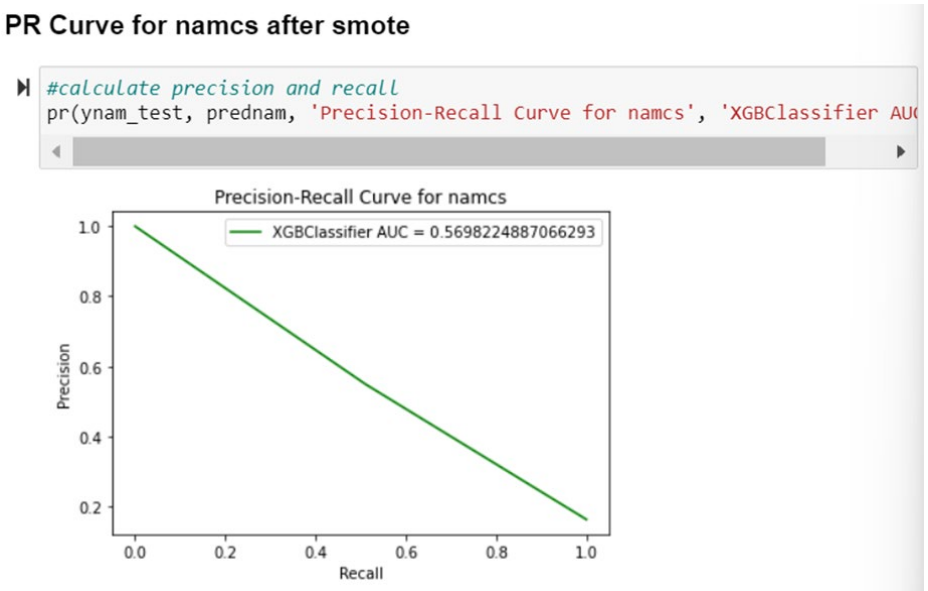

Fig. 23. PR Curve NAMCS after Smote

ROC Curve. The ROC curve for the XGBOOST classifier performed on the NAMCS dataset after SMOTE, AUC was 0.83

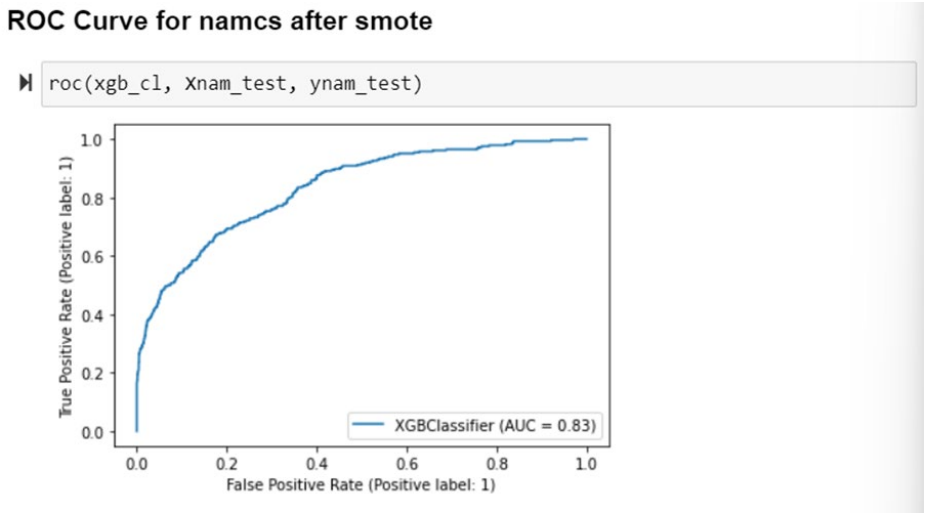

Fig. 24. ROC Curve XGBOOST

NHANES. The SMOTE function was performed to the NHANES dataset to balance the data. The XGBOOST classifier was performed by splitting the data into training and testing data. The cross-validation score of the NHANES dataset after SMOTE with XGBOOST classifier using F1_macro was 0.79 with a standard deviation of 0.17. There was a significant difference for the F1 score before and after using the SMOTE.

Model scores for nhanes after smote

```
In [46]: skmets(ynh_test, prednh)

Out[46]: {'Accuracy': 0.7380546075085325,
          'Precision': 0.4489795918367347,
          'Sensitivity_recall': 0.30662020905923343,
          'Specificity': 0.8779661016949153,
          'F1_score': 0.36438923395445133}
```

Fig. 25. NAMCS XGBOOST Performance

XGBOOST performance scores for NHANES dataset after SMOTE were:

- Accuracy: 73.8%
- Precision: 44.9%
- Sensitivity recall: 30.66%
- Specificity: 87.8%
- F1_score: 0.3644

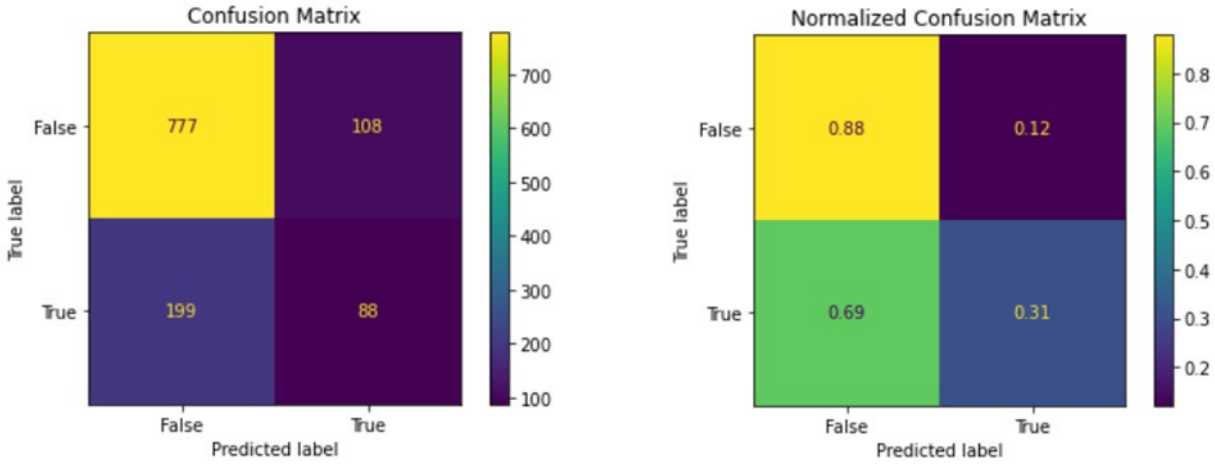

Fig. 26. Confusion Matrix of NHANES

Confusion Matrix of XGBOOST classifier. The confusion matrix then provided the model predictions for the NHANES data. Of all the positive cases from the dataset, the classifier predicted 31% as True Positives and 69% of as False Negatives. Of the

negatives for the depression from the dataset, the classifier predicted 88% as True Negatives and with 12% as False Positives.

After using the function SMOTE, the top three features that contributed to the reporting of depression in the NHANES dataset were:

- Private Insurance
- arthritis
- overweight

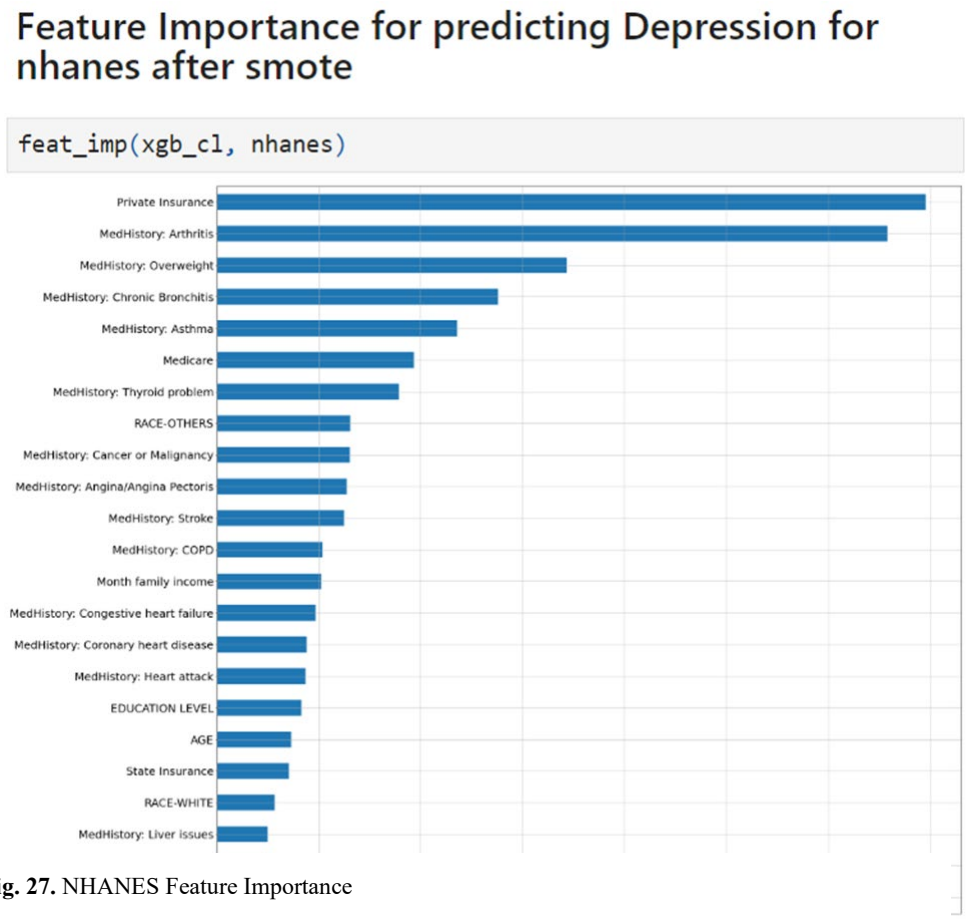

Fig. 27. NHANES Feature Importance

PR curve. The precision- recall curve for the XGBOOST classifier performed on the NHANES dataset after applying SMOTE, AUC was 0.4627

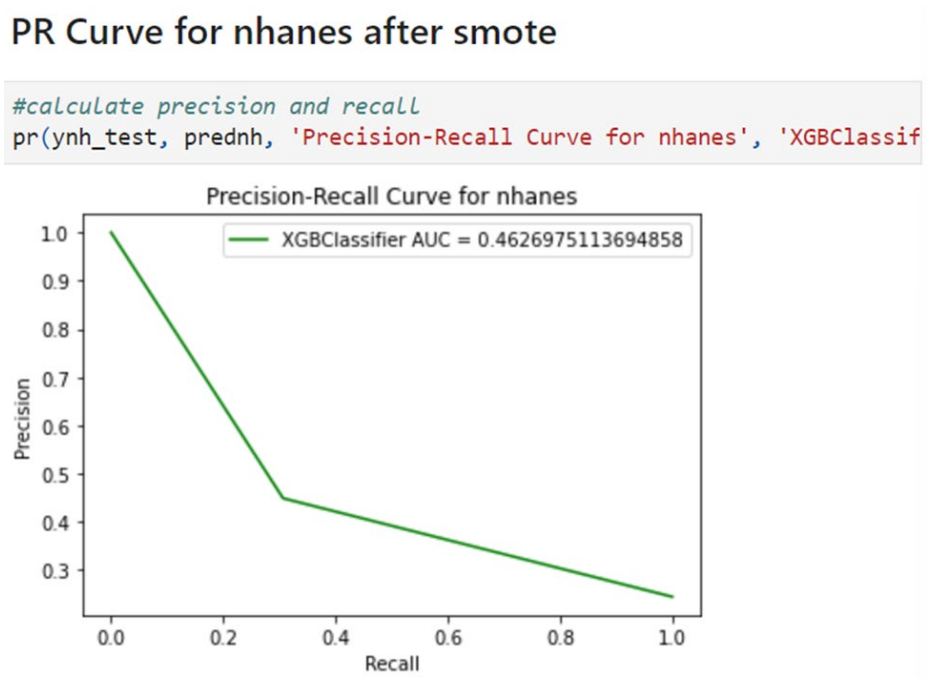

Fig. 28. PR-Curve NHANES after Smote

ROC Curve. The ROC curve for the XGBOOST classifier performed on the NHANES dataset after SMOTE, AUC was 0.65

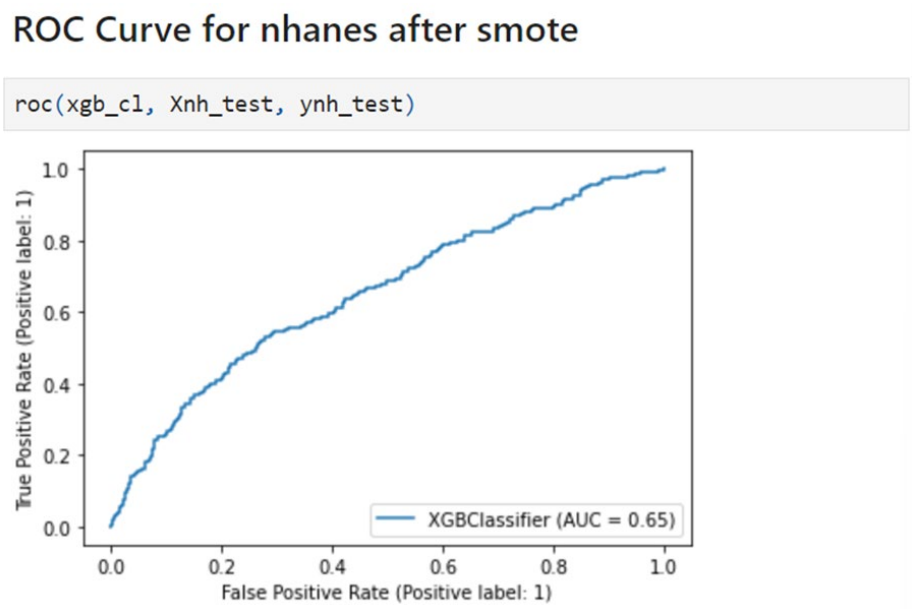

Fig. 29. ROC Curve XGBOOST

3.6 Random Forest Classifier

After performing the XGBOOST classifier on both datasets, the Random Forest Classifier was employed for classification on both datasets. The hyperparameters were fine-tuned using the Randomized Search CV.

Optimal parameters:

- n_estimators = 89,
- min_samples_split = 12,
- max_features = 'auto',
- max_depth = 28,
- bootstrap = True

NAMCS. The Random Forest Classifier was performed on the NAMCS dataset by dividing it into train and test data. The SMOTE function was also applied to this classifier to balance the data. The cross-validation score (f1_macro) of Random Forest Classifier for the NAMCS dataset was 0.87 with a standard deviation of 0.08.

```
Mean f1_macro of 0.63 with a standard deviation of 0.02


Mean f1_macro of 0.87 with a standard deviation of 0.08
```

Fig. 30. Before and After Smote NAMCS

The model performance scores of Random Forest Classifier for the NAMCS dataset were:

- Accuracy: 84.55%
- Precision: 52.59%
- Sensitivity recall: 53.52%
- Specificity: 90.6%
- F1_score: 0.53

Model scores for namcs after smote

```
skmets(ynam_test, prednam)
```

```
9]: {'Accuracy': 0.8454910970706491,
     'Precision': 0.5259515570934256,
     'Sensitivity_recall': 0.5352112676056338,
     'Specificity': 0.905971173644475,
     'F1_score': 0.5305410122164049}
```

Fig. 31. NAMCS Random Forest Performance

After performing the feature importance function on the NAMCS dataset, the top three features that contributed for the reporting of depression were:

- No Chronic Conditions
- Age
- Specialty: Surgical Care

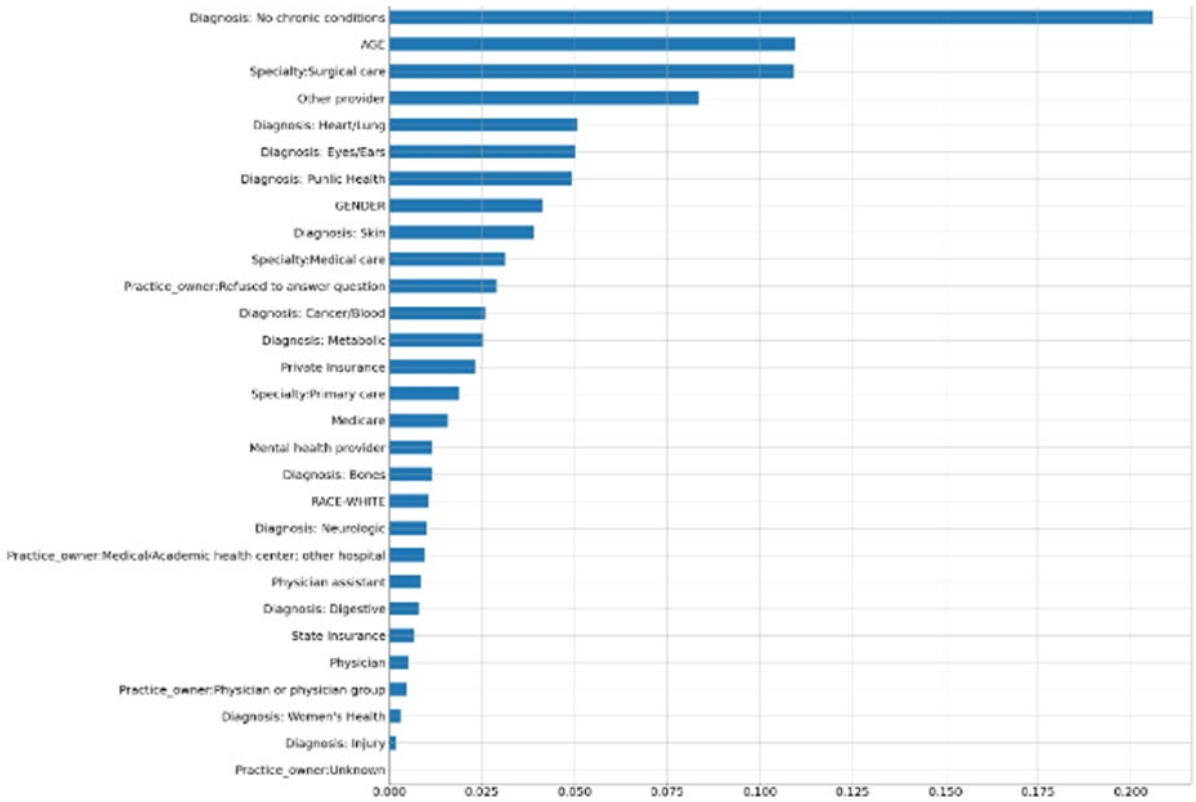

Fig. 32. NAMCS Feature Importance

Confusion Matrix of Random Forest Classifier. Model predictions were provided by the confusion matrix for the NAMCS data. Of all the positive cases from the dataset, the classifier predicted 54% as True Positives and 46% as False Negatives. Of all the negatives for depression from the dataset, the classifier predicted 90.6% as True Negatives and 9.4% as False Positives. The model has high accuracy for predicting patients with no depression.

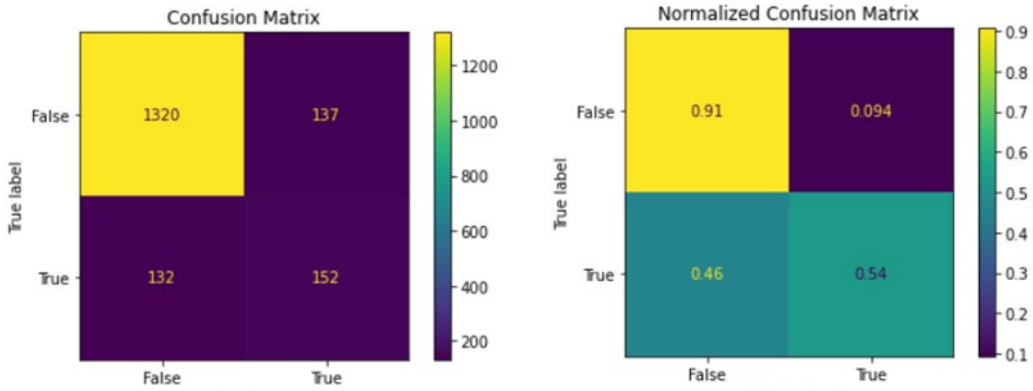

Fig. 33. Confusion Matrix for NAMCS

PR curve. The precision- recall curve for the Random Forest classifier performed on the NAMCS dataset after applying SMOTE, AUC was 0.5685

PR Curve for namcs after smote

```
#calculate precision and recall
pr(ynam_test, prednam, 'Precision-Recall Curve for namcs', 'RandomForestClas
```

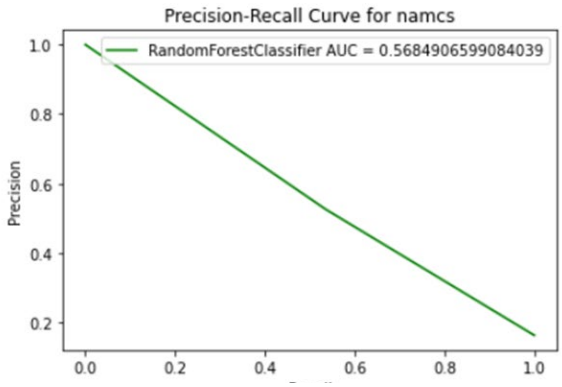

Fig. 34. PR-Curve NAMCS after Smote

ROC Curve. The ROC curve for the Random Forest classifier performed on the NAMCS dataset after SMOTE, AUC was 0.85

ROC Curve for namcs after smote

```
roc(clf, Xnam_test, ynam_test)
```

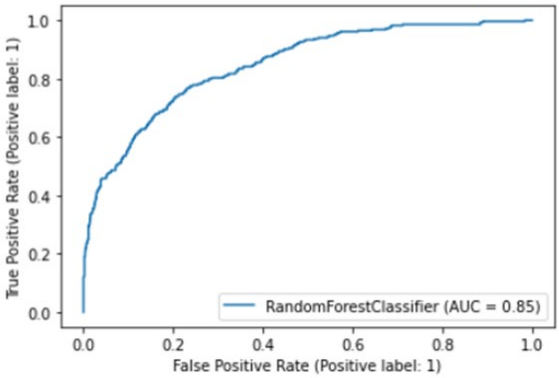

Fig. 35. ROC Curve Random Forest

NHANES. The Random Forest Classifier was performed on the NHANES dataset by splitting the data into training and testing data. SMOTE was employed to balance the data in the dataset. The cross-validation score (f1_score) of the Random Forest Classifier for the NHANES dataset was 0.79 with a standard deviation of 0.15.

```
Mean f1_macro of 0.54 with a standard deviation of 0.00

Mean f1_macro of 0.79 with a standard deviation of 0.15
```

Fig. 36. Before and After Smote NHANES

The model performance score of Random Forest Classifier for NHANES dataset were:

- Accuracy: 72.87%
- Precision: 42.36%
- Sensitivity recall: 29.97%
- Specificity: 86.78%
- F1_score: 0.3510

Model scores for nhanes after smote

```
skmets(ynh_test, prednh)

1]: {'Accuracy': 0.7286689419795221,
     'Precision': 0.4236453201970443,
     'Sensitivity_recall': 0.29965156794425085,
     'Specificity': 0.8677966101694915,
     'F1_score': 0.3510204081632653}
```

Fig. 37. NHANES Random Forest Performance

After performing the feature importance function on the NHANES dataset, the top three features that contributed to the reporting of depression were:

- Overweight
- Private Insurance
- Arthritis

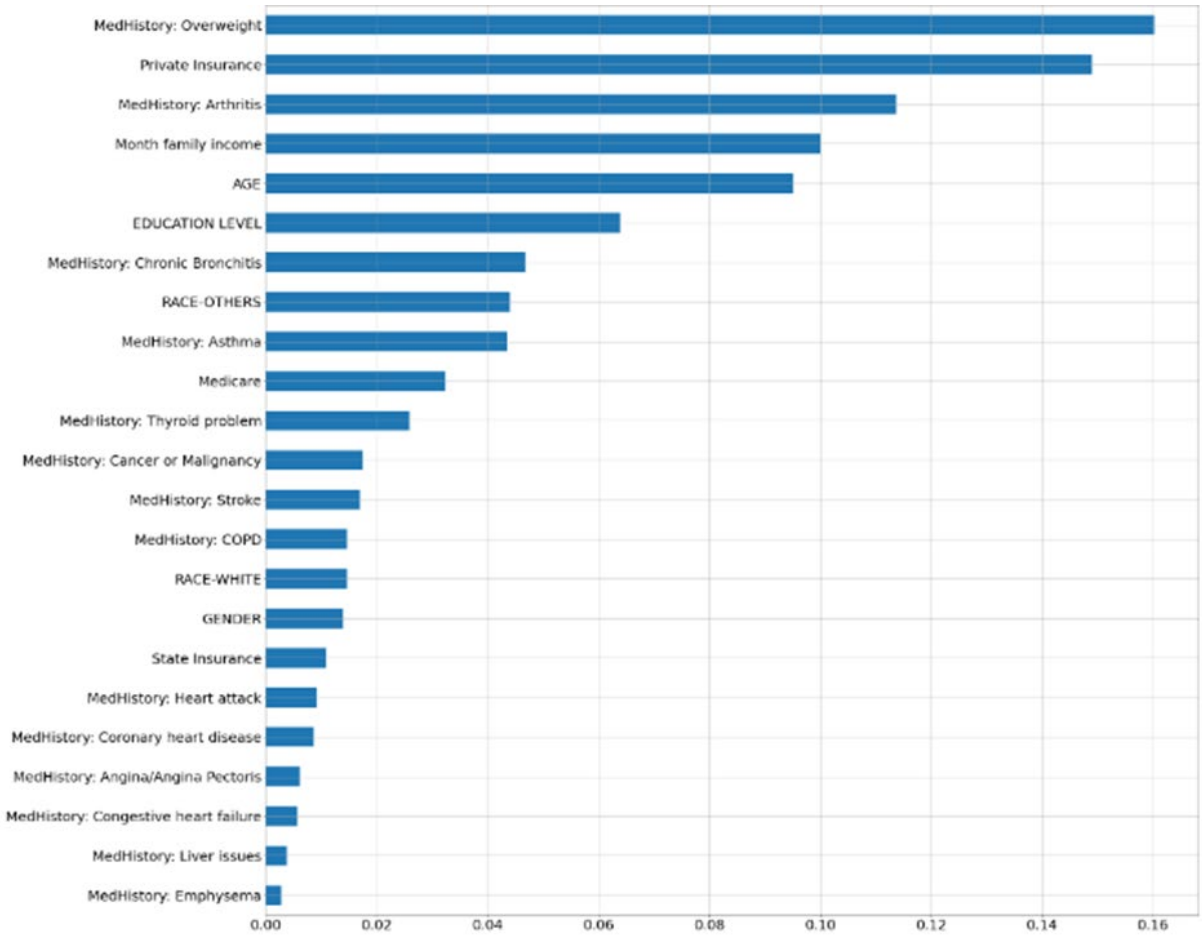

Fig. 38. NHANES Feature Importance

Confusion Matrix of Random Forest Classifier. Model predictions were provided by the confusion matrix for the NHANES data. Of the positive cases from the dataset, the classifier predicted 30% of them as True Positives and 70% of them as False Negatives. Of the negatives for depression from the dataset, the classifier predicted 87% of them as True Negatives and 13% as False Positives. The model has high accuracy for predicting patients with no depression.

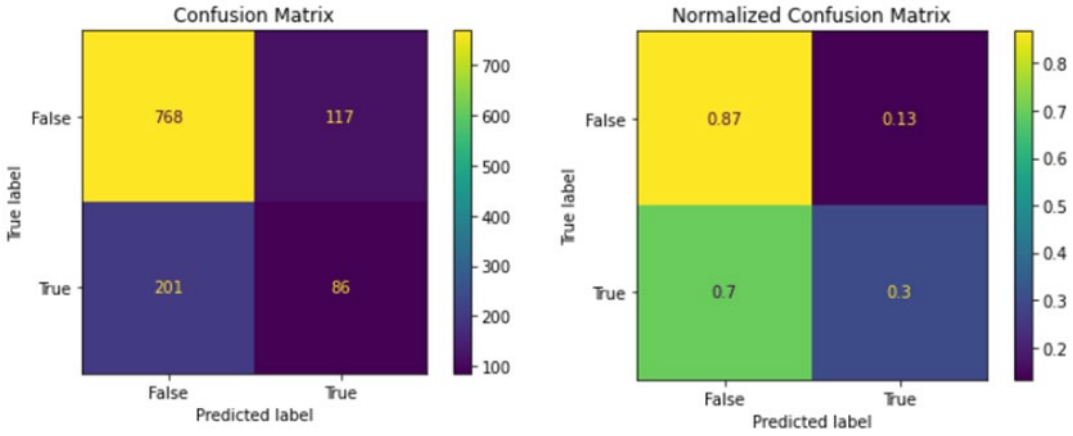

Fig. 39. Confusion Matrix for NHANES

PR curve. The precision- recall curve for the Random Forest classifier performed on the NHANES dataset after applying SMOTE, AUC was 0.4474

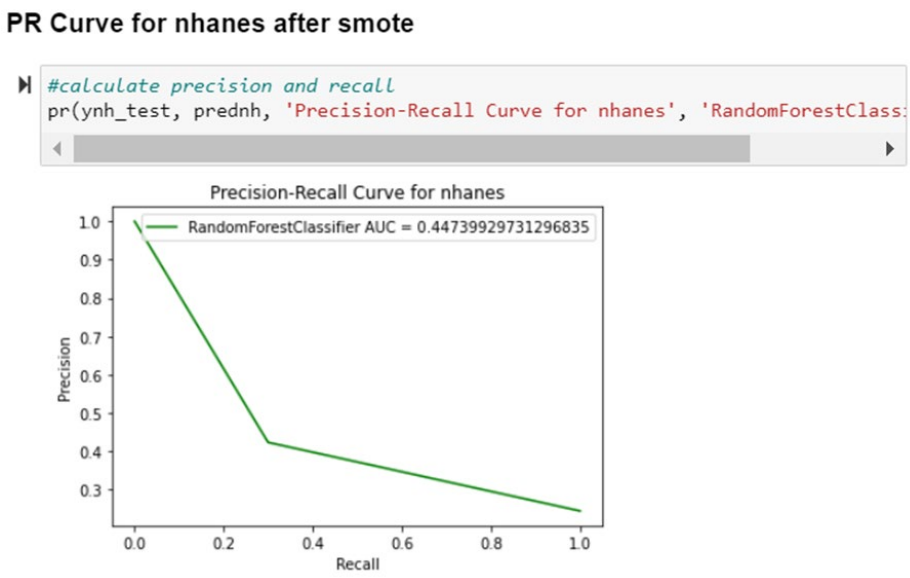

Fig. 40. PR-Curve NHANES after Smote

ROC Curve. The ROC curve for the Random Forest classifier performed on the NHANES dataset after SMOTE, AUC was 0.68

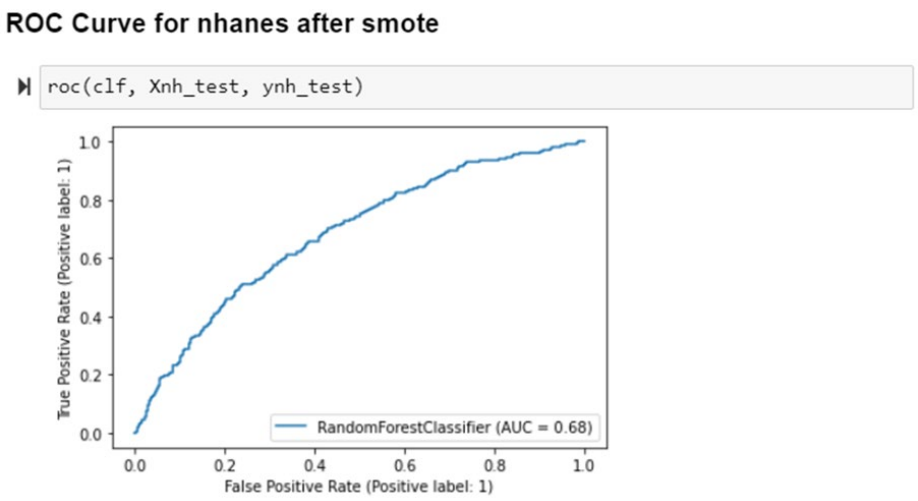

Fig. 41. ROC Curve Random Forest

4 Results

Hypothesis. An increase in the depression proportion in the NHANES dataset as compared to NAMCS dataset was found. Hypothesis testing supported the rejection of the null hypothesis with 9.1% difference in Depression Indicator scores in the NHANES dataset compared to the NAMCS dataset.

Chi-Square testing results demonstrated significant associations between the Depression Indicator and multiple factors that were also found to forecast depression using XGBOOST and Random Forest Classification. The NAMCS data was able to provide better classification for depression as compared to NHANES survey data. A person in the NAMCS dataset had a higher chance of being detected for depression compared to a person who was surveyed in NHANES.

Comorbid conditions. The diagnosis codes and comorbid conditions provided a large amount of input to the model for the Depression Indicator. Groups with the largest correlation to depression included those with conditions of:

- Overweight
- Arthritis
- Cardiopulmonary disease or Asthma
- Dermatologic disorders
- Public health conditions

The comorbid features found to have the most feature importance in predicting depression included:

- Overweight
- Arthritis.
- Cardiopulmonary
- Cancer or blood related diseases
- No Chronic Disease

The no chronic disease feature was found to be important for both predictive models for the NAMCS dataset. Although chronic disease can contribute to depression, it is important to screen those with no chronic disease as other factors may be a contributor.

Age. Age was found to be associated with the Depression Indicator in chi-square testing and as a predictor by the machine learning model in the NAMCS dataset. However, the samples of both datasets were distributed toward the older age group. This limits the findings since the datasets do not represent all age groups.

Private Insurance. Both models found private insurance to be a predictor of depression in the NHANES dataset, but the feature was of lower importance in the NAMCS dataset. Further testing may be needed to validate the correlation.

5 Conclusions

This study provides preliminary evidence for developing a machine learning program to predict the classification of depression in adults. Findings from this study can be used in future study to develop a more precise screening tool for depression. Further study is recommended based on these findings for the development and testing of a screening algorithm for providers to increase reporting from laypersons to providers. Although there were discrepancies in depression reporting, additional research is required to investigate the influence of this reporting on adults and the steps that should be taken with specific groups for better screening and diagnosis.

6 Limitations

Although every effort was made to minimize limitations, there inevitably are limitations in any research study. The use of 2018 survey data for both datasets due to constraints from the Covid pandemic's ban on in-person contact may limit the relevancy of the findings. The data is relatively recent data, however data changes rapidly in all data science, especially medical data. It is likely that depression rates and factors leading to depression have been altered from the compounding effect of the Covid pandemic on depression.

To compare the datasets, model predictions were performed for the diagnosis of depression on the two datasets. However, findings from the models related to direct comparisons of the two datasets are limited because the two samples were from two separate populations.

The sample populations were not normally distributed for age. Findings cannot be generalized to all age groups since the samples were skewed toward the older population.

7 Project Challenges and Successes

Initially, the team was meant to do a comparative analysis by identifying the common variables in both datasets and analyze the mean differences in depression reporting by providers compared to the lay public. The project intended to identify any discrepancies in the identification of patients with depression by providers and find missed diagnoses of depression. Nonetheless, after consultation with our advisor and further investigation into the data, the populations were not the same and cannot be compared directly. The team shifted away from investigating for those in the NAMCS group similar to the NHANES, but not diagnosed with depression to comparing the two datasets for similarities in sample demographics and comparison of Depression Indicators.

Also, after getting full access to the data, the team realized that most of the variables were categorical, and t-testing was not appropriate for hypothesis testing. To ob-

tain pertinent findings, chi-square analyses were required to identify significant variables, and modeling was possible to conclude the hypothesis testing.

The team was meticulous about details and had to rerun the codes several times to attain the required results. After cleaning and running the codes, we saw anomalies in our methods and considered several approaches to ensure we were doing the project correctly and on time. Choosing two large datasets increased the complexity of the project and made it difficult to complete all analyses and interpretations on time. Two datasets also increased the complexity of the data interpretations.

8 Appendix

CONTINUOUS DATA										
DATASET	n	VARIABLE	MEAN	MEDIAN	MODE	STD	VARIANCE	IQR	SKEWNESS	KURTOSIS
NAMCS	8701	Age	58	61	71	18.44	340.1	27	0	-0.74
NHANES	5856	Age	50	51	80	18.78	352.54	32	-0.06	-1.17
NHANES	5856	PHQ Score	2.82	1	0	4.11	16.89	4	2.08	4.67

Fig. 42. Descriptive Statistics-Continuous Variables

CATEGORICAL DATA						
	NHANES					
VARIABLE	MODE	STD	VARI	IQR	SKEWNESS	KURTOSIS
Gender	1	0.50	0.25	1	-0.06	-2.00
Black	0	0.42	0.18	0	1.29	-0.34
Other Race	0	0.49	0.24	1	0.31	-1.91
White	0	0.48	0.23	1	0.64	-1.59
Education	4	1.55	2.40	3	-0.44	-1.33
Income	15	4.98	24.81	10	0.04	-1.17
Medicare	0	0.44	0.20	1	1.04	-0.91
Uninsured	0	0.01	0.00	0	76.52	5856.00
Commercial	0	0.50	0.25	1	0.05	-2.00
State-funded	0	0.38	0.14	0	1.71	0.93
Asthma	0	0.36	0.13	0	1.95	1.79
Overweight	0	0.49	0.24	1	0.49	-1.76
Arthritis	0	0.45	0.21	1	0.93	-1.14
CHF	0	0.18	0.03	0	5.12	24.19
Vascular	0	0.21	0.04	0	4.38	17.16
Chest Pain	0	0.16	0.03	0	5.78	31.43
Heart attack	0	0.21	0.04	0	4.33	16.75
Stroke	0	0.21	0.04	0	4.30	16.51
Emphysema	0	0.13	0.02	0	7.23	50.31
Bronchitis	0	0.25	0.06	0	3.45	9.91
Thyroid	0	0.32	0.10	0	2.46	4.07
COPD	0	0.22	0.05	0	4.13	15.05
Cancer	0	0.30	0.09	0	2.66	5.08
Liver	0	0.22	0.05	0	4.12	14.99
DPQ1	0	0.72	0.52	0	2.31	4.74
DPQ2	0	0.68	0.46	0	2.48	5.84
DPQ3	0	0.91	0.83	1	1.60	1.44
DPQ4	0	0.91	0.83	1	1.33	0.83
DPQ5	0	0.74	0.55	0	2.32	4.68
DPQ6	0	0.59	0.34	0	3.18	10.30
DPQ7	0	0.64	0.41	0	3.10	9.27
DPQ8	0	0.51	0.26	0	4.01	16.56
DPQ9	0	0.28	0.08	0	7.38	61.53
PHQ Score	0	4.11	16.89	4	2.08	4.67
Feels DPRN	5	1.27	1.61	1	-1.40	1.19
DPRN Meds	2	0.39	0.15	0	-2.99	8.73
DPRN Level	0	1.15	1.31	2	0.80	-0.86
DI	0	0.43	0.19	0	1.18	-0.60

	NAMCS					
VARIABLE	MODE	STD	VARIANCE	IQR	SKEWNESS	KURTOSIS
Gender	0	0.49	0.24	1	0.31	-1.90
Black	0	0.26	0.07	0	3.25	8.60
Other Race	0	0.23	0.05	0	3.77	12.20
White	1	0.34	0.11	0	-2.17	2.71
Speciality Care	0	0.46	0.21	1	0.84	-1.29
Primary Care	0	0.44	0.20	1	1.05	-0.91
Surgery Care	0	0.49	0.24	1	0.30	-1.91
Nurse	0	0.30	0.09	0	2.65	5.05
Provider-withheld	0	0.03	0.00	0	31.05	962.33
NPNMW	0	0.11	0.01	0	8.98	78.72
MHP	0	0.11	0.01	0	9.22	82.95
Other provider	0	0.46	0.21	1	0.83	-1.31
MD/DO	1	0.11	0.01	0	-8.49	70.06
PA	0	0.18	0.03	0	5.12	24.25
No provider	0	0.02	0.00	0	46.62	2171.50
Hospital Owner	0	0.21	0.05	0	4.21	15.73
Insurance Owner	0	0.32	0.10	0	2.36	3.57
Phys.Owner	1	0.41	0.17	0	-1.41	-0.02
Owner-withheld	0	0.06	0.00	0	17.87	317.45
Medicare	0	0.48	0.23	1	0.60	-1.64
Uninsured	0	0.15	0.02	0	6.17	36.09
Commercial	0	0.50	0.25	1	0.26	-1.93
State-funded	0	0.27	0.07	0	3.09	7.58
Infection	0	0.04	0.00	0	23.26	539.12
Urinary	0	0.23	0.05	0	3.88	13.08
Womens Health	0	0.12	0.01	0	8.10	63.56
Genetic	0	0.05	0.00	0	19.82	390.73
Other Disorder	0	0.28	0.08	0	3.01	7.03
Injury	0	0.15	0.02	0	6.38	38.77
Morbidity	0	0.00	0.00	0	0.00	0.00
Public health	0	0.37	0.13	0	1.87	1.50
Cancer/Blood	0	0.18	0.03	0	5.16	24.65
Metabolic	0	0.21	0.05	0	4.27	16.23
Neurologic	0	0.17	0.03	0	5.49	28.13
Eyes/Ears	0	0.35	0.12	0	2.07	2.30
Heart/Lung	0	0.26	0.07	0	3.20	8.27
Digestive	0	0.17	0.03	0	5.37	26.80
Skin	0	0.28	0.08	0	2.89	6.37
Bones	0	0.25	0.06	0	3.47	10.05
No illness	0	0.47	0.22	1	0.70	-1.50
Chronic DPRN	0	0.30	0.09	0	2.61	4.81
Therapy Referral	0	0.14	0.02	0	6.97	46.53
MHP Referral	0	0.14	0.02	0	6.76	43.66
DPRN Screen	0	0.20	0.04	0	4.58	19.02
DI	0	0.36	0.13	0	1.91	1.66

Fig. 43. Descriptive Statistics-Categorical Variables

9 References

1. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health
2. World Health Organization: Mental Health and COVID-19: Early evidence of the pandemic's impact (2022)
3. https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/
4. Knaak, S., Mantler, E., Szeto, A.: Mental illness-related stigma in healthcare. Healthcare Management Forum 30, 111-116 (2017)
5. Hansson, L., Jormfeldt, H., Svedberg, P., Svensson, B.: Mental health professionals' attitudes towards people with mental illness: do they differ from attitudes held by people with mental illness? Int J Soc Psychiatry 59, 48-54 (2013)
6. Garcia, M.E., Hinton, L., Neuhaus, J., Feldman, M., Livaudais-Toman, J., Karliner, L.S.: Equitability of Depression Screening After Implementation of General Adult Screening in Primary Care. JAMA Network Open 5, e2227658-e2227658 (2022)
7. National Center for Health Statistics: Summary of Current Surveys and Data Collection Systems. factsheet, Centers for Disease Control and Prevention (2020)
8. National Center for Health Statistics (NCHS): National Health and Nutrition Examination Survey. In: Centers for Disease Control and Prevention (CDC) (ed.). U.S. Department of Health and Human Services, Centers for Disease Control and Prevention,, Hyattsville, MD (2018)
9. National Center for Health Statistics: Unweighted Response Rates for NHANES 2017-2018 by Age and Gender. In: NHANES-2017-2018-Response-Rates-508 (ed.). National Center for Health Statistics,, https://wwwn.cdc.gov/nchs/data/nhanes3/ResponseRates/NHANES-2017-2018-Response-Rates-508.pdf
10. National Center for Health Statistics: National Ambulatory Medical Care Survey. Center for Disease Control, (2018)
11. National Center for Health Statistics: National Ambulatory Medical Care Survey: 2018 National Summary Tables. Centers for Disease Control and Prevention, (2018)
12. National Center for Health Statistics (NCHS): 2018 NAMCS Micro-Data File Documentation. In: Ambulatory and Hospital Care Statistics (ed.), pp. 155. Division of Health Care Statistics, (2018)
13. https://wwwn.cdc.gov/nchs/nhanes/tutorials/module1.aspx