Project Report on Mental Health Analysis

By Neha Shetty, Shubhangi Kishore, Melita Saldanha, and Ronak Khandelwal

Introduction:

The 2030 Agenda for Sustainable Development adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future [1]. At its heart are the 17 Sustainable Development Goals (SDGs), one of which is ensuring 'Good Health and Well-being'. In this project we attempted to solve this SDG by analyzing Mental Health which is one of the aspects contributing to a healthy life. Mental health conditions cause a great deal of distress or impairment. The most frequent mental disorder is Depression, which affected approximately 7.1% of the adults in the United States in the past year, and yet not everyone is comfortable talking about it. The unmet need for treatment of mental disorders is a major problem. The ongoing pandemic only worsens this situation and hence the need to analyze this issue is important.

We considered the NHANES dataset which spans over 20 years of survey data and helps identify patterns within the data that would otherwise be hard to detect. Identifying factors such as biological causes, demographic sectors, and the lifestyle of an individual, would help in understanding the symptoms of depression and lead to a faster and accurate diagnosis and treatment plan. Such a vast collection of data available to gain insights from, required the use of Big Data frameworks for its analysis. Also, since surveys are conducted and reviewed yearly, the data available is inconsistent across the years span and hence, constructing a concrete dataset from survey data is difficult. Due to design considerations, variables may be added, removed or renamed for a new survey year, or the format might be changed totally. Therefore, compiling a complete dataset involves manually searching for variables across yearly releases, and adds additional steps besides the regular pre-processing of data.

Background:

While there are factors that cause Depression, there are also factors that are the consequence of an individual suffering from it. Many people fail to notice when a friend or family member is on the verge of falling into depression. By analyzing data about individuals suffering from depression, we can find the factors which cause and are the consequences of depression. This will help to take preventive measures for an individual who is on the verge of falling into depression. Also, this study will help people who are already depressed to be diagnosed sooner and be provided with the right type of help.

Several studies have been done in this domain. The data considered is usually from social media statistics. Also, a small subset of features from specific domains related to mental health have been considered or the data considered is for a small time frame. We approached this problem by considering a vast number of features, some of which may not generally be directly associated with mental health. There is a possibility that some factor does not significantly affect the mental health of an individual independently, but when combined with other factors, it tends to have a higher contribution to the downfall of an individual's mental health. Analysis of such correlations in the various factors contributing to one's mental health requires a large dataset of individuals suffering from depression and hence, requires Big Data frameworks for its computation.

Data:

We used the National Health and Nutrition Examination Survey (NHANES) dataset [2]. NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the

United States. The interview includes demographic, socioeconomic, dietary, and health-related questions in addition to the laboratory tests administered by highly trained medical personnel.

The study focuses on questionnaire data and demographics to get details about individual's mental health, alcoholism, drug usage. Features were selected across views including laboratory and examination using correlation scores. We had 149 datasets with around 10-48 questions in each of them. We merged these datasets on one common column that is the Sequence number unique to everyone interviewed. The dataset spans over 20 years and has 6000 unique observation points for each year with over 100 attributes for each row.

The data from the questionnaire dataset was mainly categorical while the data from other datasets were continuous values. Each of the dataset came with the description for each of the questions asked which made it easy for us to understand the values better.

Features	Some of the questions considered for our analysis
Alcohol Use	1. Average number of alcoholic drinks per day. 2. How often drink alcohol over past year?
Housing Characteristics	1. Home owned, bought, rented or other? 2. Water treatment devices used or not? 3. Source of tap water?
Income	1. Monthly Family Income 2. Family Monthly poverty index
Demographics	1. Gender 2. Age 3. Race/Ethnicity 4. Marital Status
Laboratory/Examination	1. Cholesterol 2. Blood pressure 3. Oral Health

Table 1: Some of the data features and questions considered for analysis

Methods:

In the current study, we develop methods for estimating the subjective well-being, calculated as the depression score against survey-based ground truth measures of the NHANES Dataset. To understand this relationship we perform three types of statistical analyses: (1) Correlation analysis (2) Mediation analysis (3) Prediction

Data preprocessing

In this project, each of these categories (demographics, laboratory tests) are thought of as separate views of the participants' data. The datasets were merged as required using the unique identifier of the interviewee. The features obtained from applying multiview techniques to the data optimize properties that help to incorporate information from the different views into a single representation.
Categorical values were encoded to integers and scaled to understand the data better. The sample weights used followed the National Center for Health Statistics guidelines and were calculated according to the base probabilities of selection, adjusted for nonresponse, to match population control totals. Missing values were handled by removing or imputing by reasoning why the values might be missing.

Empirical Data Analysis

With the preprocessed data, we did some Empirical Data Analysis to understand the data better and to see if this data really helps us to achieve our goal. We visualize significant correlations by controlling for

age, gender and sexual orientation. **Principal Component Analysis (PCA)** was also used for exploratory data analysis. Observations suggest single and separated people show more signs of depression as compared to married or widowed. Moreover, females who are separated from their partner exhibit more signs of depression. Controlling for age groups, mid aged Non-Hispanic Black people and old aged Non-Hispanic White people show more signs of depression. We formulated hypothesis and used **T test** for evaluation, adjusting for multiple comparisons by applying a **Benjamini—Hochberg false discovery rate correction** to the significance threshold ($p < .05$).

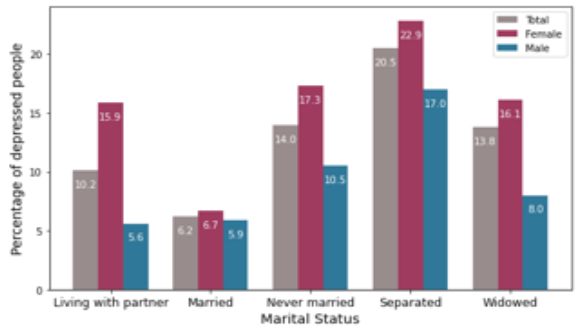


Figure 1: Marital Status of each gender vs Depression level

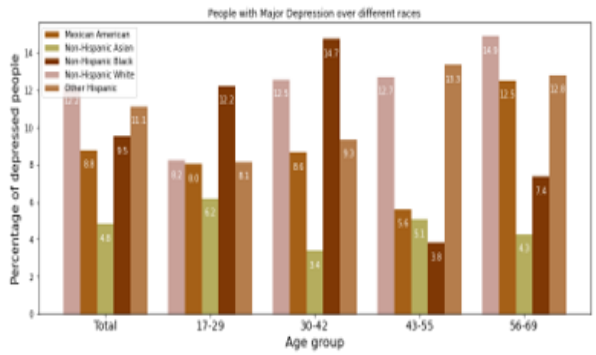


Figure 2: Ethnicity of each age group vs Depression level

Mediation Analysis

Mediation analysis tests a hypothetical causal chain and helps better understand the relationship between dependent and independent variables that show higher correlation by introducing a mediator. A mediating variable transmits the effect of an independent variable on a dependent variable. We follow the standard three-step, Baron and Kenny approach - 1: we regress our independent and dependent variable in a standard OLS regression. 2: we regress the independent variable and mediator. 3 we create a multivariate model and regress both the mediator and independent variable with depression score.

Feature Engineering

We used a feature selection pipeline on our data. After preprocessing the data, we perform **Linear regression** for each of the features on 149 datasets (with the number of features ranging from 10 to 48) on a 20 year pooled survey data using **PySpark** and **HDFS**. Loading and processing the multistage NHANES dataset using the **Spark** framework enabled the calculation of test statistics and beta values by performing linear regression. The top 50 beta values allows us to select most correlated features and most relevant datasets for the prediction task. Multivariate LR is performed on the Questionnaire dataset using **Tensorflow** with the following hyperparameters (epochs - 10, learning rate - 0.001, momentum - 0.9).

Prediction

The features selected were found to be robust both in terms of interpretability and predictive power. The target variable was depression scores, calculated as a weighted average of individual level response to the depression questionnaire in the NHANES dataset. The algorithm used for prediction is **Nearest Neighbour** where the number of neighbors was estimated using the elbow technique. Cosine similarity was used as the similarity measure. Predictive accuracy is measured in terms of MSE between the actual values and the predicted values, whereas standard errors are calculated across 10 folds.

Evaluation and Results:

In this section we first try to show the results of the hypothesis testing which allowed us to filter a few features from 1000's of features available for predictions. Finally, we look at the results of the recommendation engine delineating the similarity scores, predicted values and MSE.

Hypothesis testing

We begin by looking at correlations between depression scores and socioeconomic features and known risk factors. With access to the cleaned and combined dataset we stated hypotheses and tested for their correctness. We began with the ones that are presumed to have a relationship with mental health, and then moved onto novel features that indicated correlation obtained through **LR**. We tested numerous hypotheses. As evident, a number of our hypotheses resulted in rejecting decisions. For some, like the quantity of the toothpaste used, as accepted but others like BMI and vitamin deficiency were surprises.

Hypothesis	Results
Relationship status affects mental health of an individual	Failed to reject
BMI of an individual affects the mental health	Rejected
Individuals less severely affected by mental health problems have higher incomes	Rejected
Alcohol abuse has strong correlation with mental health issues	Rejected
People with vitamin deficiency are known to have mental health issues	Rejected
People who rated their dental health better show fewer signs of depression	Failed to reject
Home ownership is positively correlated with Mental health of an individual	Failed to reject
Severity of mental health issues show a negative correlation with expenditure of food	Rejected

Table 2: Output of Hypothesis Testing

We'll now take a look at how some of the features performed in the testing. Figure 3 shows the relationships between different depression groups and the money spent on food. As visible there is no strong correlation between the two. Most of the other novel features like the amount of toothpaste used daily and presence of dairy in diet shows similar trends. Further, figure 4 shows the feature importance of some of the features which show the most promise. The plot shows relative importance of the features like age, sexual preference, gender etc. We can clearly see that the age of an individual(RIDAGEYR) certainly has an effect on his/her mental health and so does the monthly family income(INDFMIN2).

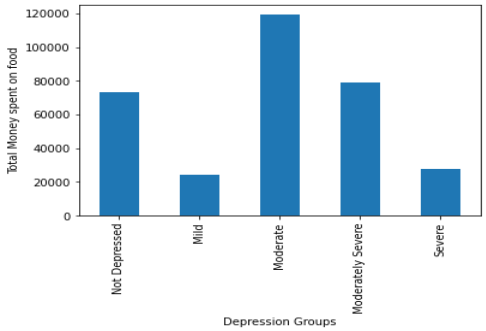
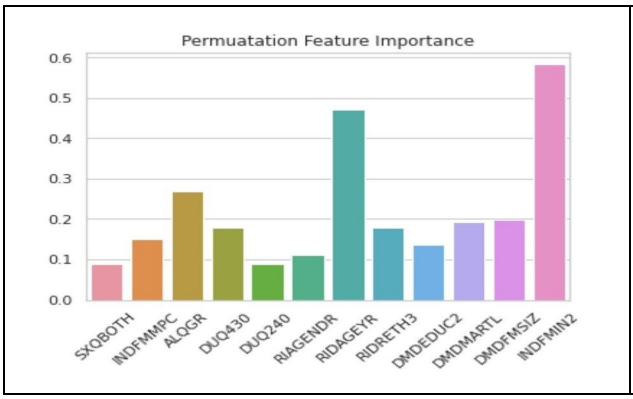


Figure 3: Money spent on food vs depression groups



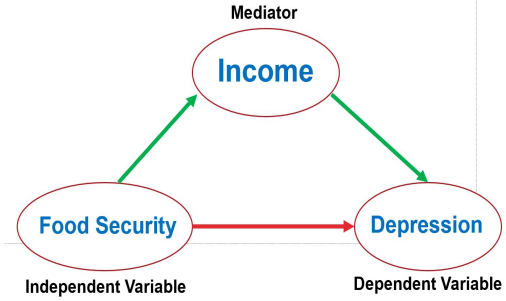
Feature Importance ordered by rank -

Annual Income
Age Group
Alcohot
Family Size
Marital Status
Drug Use
Race

Figure 4: Output from hypothesis testing

Mediation Analysis

The table below shows the results from mediation analysis of our target variable depression score with the predictor being food security. With food security alone the relationship b/w the two variables were not statistically clear, introducing the feature income as a mediator got better results as shown in Figure 5. In our study we observed indirect effects of food security and depression by using income as a mediator. We also observed statistically higher correlation between Alcoholism and depression using rehabilitation as a mediator.



	Estimate	Lower CI bound	Upper CI bound
ACME (control)	0.174327	0.164632	0.184200
ACME (treated)	0.144734	0.136557	0.153053
ADE (control)	-0.126808	-0.131602	-0.122037
ADE (treated)	-0.156402	-0.161304	-0.151825
Total effect	0.017925	0.008858	0.027770
Prop. mediated (control)	9.782667	6.577363	18.791141
Prop. mediated (treated)	8.100491	5.520734	15.719491
ACME (average)	0.159530	0.151188	0.168452
ADE (average)	-0.141605	-0.145047	-0.138421
Prop. mediated (average)	8.927239	6.092903	17.367857

Figure 5: Results from mediation analysis

Predictions

After carefully selecting the features for the prediction engine we used the output csv from previous sections. In the spark program by combining the different feature-sets we tried finding the neighbors for various individuals using the features showing most promise as above. After calculating the similarities, we used the weighted average approach to predict if a person is depressed or not based on the categories in the NHANES Analytics Notes. Finally, we calculate the **MSE** from the predicted values. This particular value of MSE shows that we were able to successfully predict a significant portion of the dataset.

```
>>> inverted_similarities.take(10)
[(41509.0, (71682.0, 0.8997328023941468)), (41540.0, (71682.0, 0.9076895676680733)), (41546.0, (71682.0, 0.95627378
95916898)), (41589.0, (71682.0, 0.9318013318625441)), (41636.0, (71682.0, 0.9119838077477359)), (41718.0, (71682.0,
 0.9226205111720094)), (41775.0, (71682.0, 0.9314294510190376)), (41802.0, (71682.0, 0.943671837777529)), (41822.0,
 (71682.0, 0.9239044310524365)), (41864.0, (71682.0, 0.9419135777181143))]
```

Figure 6: Similarities

```
>>> rounded_preds.take(10)
[(71682.0, (0.0, 0)), (92168.0, (0.0, 0)), (61962.0, (0.0, 0)), (57052.0, (1.0, 0)), (83202.0, (0.0, 0)), (68110.0,
 (0.0, 0)), (87640.0, (0.0, 0)), (70674.0, (0.0, 0)), (74264.0, (0.0, 0)), (46170.0, (1.0, 0))]
```

Figure 7: Predictions

```
>>> MSE = rounded_preds.map(lambda r: (r[1][0] - r[1][1]) ** 2).mean()
>>> print("Mean Squared Error = " + str(MSE))
Mean Squared Error = 0.447115384615
```

Figure 8: Output from prediction engine

Conclusion:

With this project we tried to identify the relationship between lifestyle factors like dental health and consumer spending with mental health of an individual. We used the NHANES dataset with technologies like **Spark, Hadoop, TensorFlow**, and concepts like **Cosine Similarity, Correlation, Mediation analysis**, to achieve our results. We can conclude that there are indeed some lifestyle factors which provide relatively strong correlation with the mental health of the individual but since correlation does not imply causation there is a need for more detailed studies and more data for each of the factors. Our key contributions can be summarized as follows - (1) Applying standard reweighting techniques, bias correction weights on a huge multiview complex survey (2) Developed methods to evaluate depression and alcoholism and engineer features from questionnaire data (3) Performing Mediation analysis to observe indirect correlation (4) Performing correlation analysis on 4831 features and computing weights (5) Apply Nearest Neighbor techniques to obtain prediction accuracies on mental health score and depression category.

References:

[1] 2030 Agenda for Sustainable Development

[2] NHANES Dataset, Key Concepts about NHANES Data Structure

[3] Ferketich, Amy K., et al. "Depression as an antecedent to heart disease among women and men in the NHANES I study."

[4] Analysis by fedorgrab

[5]Tran, Tung, and Ramakanth Kavuluru. “Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks.” Journal of biomedical informatics.

[6] Analysis by deepai

[7] Resnik, Philip, et al. "Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter." Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology

[8] Analysis by jmc2392

[9] Dipnall, Joanna F., et al. "Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression."

[10] Coppersmith, Glen, Mark Dredze, and Craig Harman. "Quantifying mental health signals in Twitter." Proceedings of the workshop on computational linguistics and clinical psychology

[11] Miranda Olff. 2015. Mobile mental health: A challenging research agenda. European journal of psychotraumatology 6 (2015).

[12] Dipnall, Joanna F et al. “Into the Bowels of Depression: Unravelling Medical Symptoms Associated with Depression by Applying MachineLearning Techniques to a Community Based Population Sample.”