# Heinz 95-845: Nutrition, Socioeconomic Status and Depression – A Machine-Learning Approach Applied Analytics: the Machine Learning Pipeline

**Sarah Cho** jscho/jscho@andrew.cmu.edu
**Maggie Lu** yaol4/yaol4@andrew.cmu.edu
**Askari Shah** syedmehs/askari@cmu.edu
*Heinz College*
*Carnegie Mellon University*
*Pittsburgh, PA, United States*

## Abstract

Depression and mental health issues have become more pervasive in the United States. While studies linking socioeconomic status (SES) and nutrition with depression have been conducted, there are still limitations to these studies. The objective of this paper is to explore predictive methods for depression based on SES and nutrition and identify the best performing model. Algorithms such as logistic regression, decision tree, boosting, bagging, and K-nearest neighbors (KNN) are applied and results are discussed. Few external factors such as vitamin C, alcohol consumption, potassium, and vitamin B3 have been identified as having strong predictive abilities for depression.

## 1. Introduction

In recent years, the relationship between nutrition, socioeconomic status (SES) and mental health has emerged as a topic of great interest with depression and mental health issues becoming more pervasive in the United States. According to the National Institute of Mental Health (NIMH), about 6.7% of all U.S adults have suffered at least one major depressive episode in year 2016 (et al. (2017)). Earlier this year, the Center for Disease Control and Prevention (CDC) reports that suicides are becoming more often in every age group (see H Hedegaard and M Warner (2018)).

Even though there have been many studies done on various population groups in exploring the relationships between SES, diets/nutrition and mental health, we find that the majority of research has one or more of the following limitations: (1) there is very limited number of literature that targets the U.S population specifically, (2) many of these studies do not to take both SES and nutrition data as input, thus ignoring the possibility that SES and nutrition can both play a role in depression, or that SES may be a confounder for nutrition– there has been some argument on if there is a relationship between one's SES and nutrition intake (see Darmon and Drewnowski (2008); Pechey and Monsivais (2016) and Miyaki et al. (2013)), (3) most studies fail to explore statistical models other than Logistic Regression and, in some cases, Structural Equation Modeling.

In order to address these limitations, and to potentially offer new insights on this topic, we would like to reexamine how SES, and nutrition can affect one's mental health using a variety of techniques thanks to the recent advances in Machine Learning. Specifically, we use the following models in order to obtain a more comprehensive understanding on this topic: Logistic Regression, Bootstrap Aggregation (Bagging), Boosting, Random Forests,

and K-Nearest-Neighbors. We also intend to compare the accuracy and other relevant metrics of each model as well as discuss their shortcomings.

In Section 2, we offer some background information on the Machine Learning models we utilize. In Section 3, we would like to provide a detailed flowchart of our experimental setup from selecting data and features to establishing metrics for model evaluation. From there, we would go into detail on how we implement the Machine Learning models in Section 4 and discuss our findings in Section 5.

## 2. Background

### 2.1 Socioeconomic Status and Nutrition

Even though researches on the relationships between SES and the quality of diet have largely indicated a strong correlation between high SES and better diet quality, little evidence indicates that SES affects macro-nutrient composition of one's diet (Darmon and Drewnowski (2008)). Therefore for this paper, we assume that one's SES and one's macro-nutrient levels are independent.

### 2.2 Logistic Regression

Logistic regression is a predictive analytic tool where binary outcome variables is modeled with independent variables. It calculates the probability of an input being present in a class (class membership). In other words, it models the probability that an input belongs to a class (in this case, depressed or not depressed). (Dreiseitl and Ohno-Machado (2002)).

### 2.3 Decision (Classification) Tree

In this study, decision tree is used for binary classification (depressed or not depressed). Decision tree is a supervised learning model where multiple iterations of input data being split by decision nodes and outcomes are classified at the leaf nodes. (dt)

### 2.4 Bagging

Bagging, also called bootstrap aggregation, is a machine learning ensemble meta-algorithm. Bagging is designed to improve stability, accuracy and variance of classification and regression algorithms in machine learning. Bagging is particularly useful for avoiding over-fitting as it improves (reduces) variance. Bagging is generally applied to decision trees, however, it can be applied to any type of models. (Brownlee (2016)).

### 2.5 Boosting

Boosting is also a machine learning ensemble meta-algorithm. Boosting is designed to reduce variance and bias in supervised learning. Boosting combines a set of weak learners to create an overall strong learner. Boosting works on the principles of iterative learning. After each iteration, miss-classified examples are given higher weights so that the algorithm can focus on predicting them correctly in the next iteration. (D'Souza (2018)).

### 2.6 K-Nearest Neighbor

K-Nearest Neighbor is a non-parametric method used for both classification and regression. In classification tasks, the input of the K-Nearest Neighbor algorithm resides in a feature space (the vector space associated with a n-dimensional vector of numerical features) and the output is the class membership of each object which is assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small) (Altman (1992)).

## 3. Experimental Setup

The flowchart given below outlines our overall experimental design.

### 3.1 Cohort Selection

We use 2015-2016 National Health and Nutrition Examination Survey (NHANES) data. The demographic statistics are shown in Table 1.

Some of the choices we made: we eliminated all examples where depression data was given as N/A. Also, we have only considered adults (age 18+) for the purpose of our analysis.

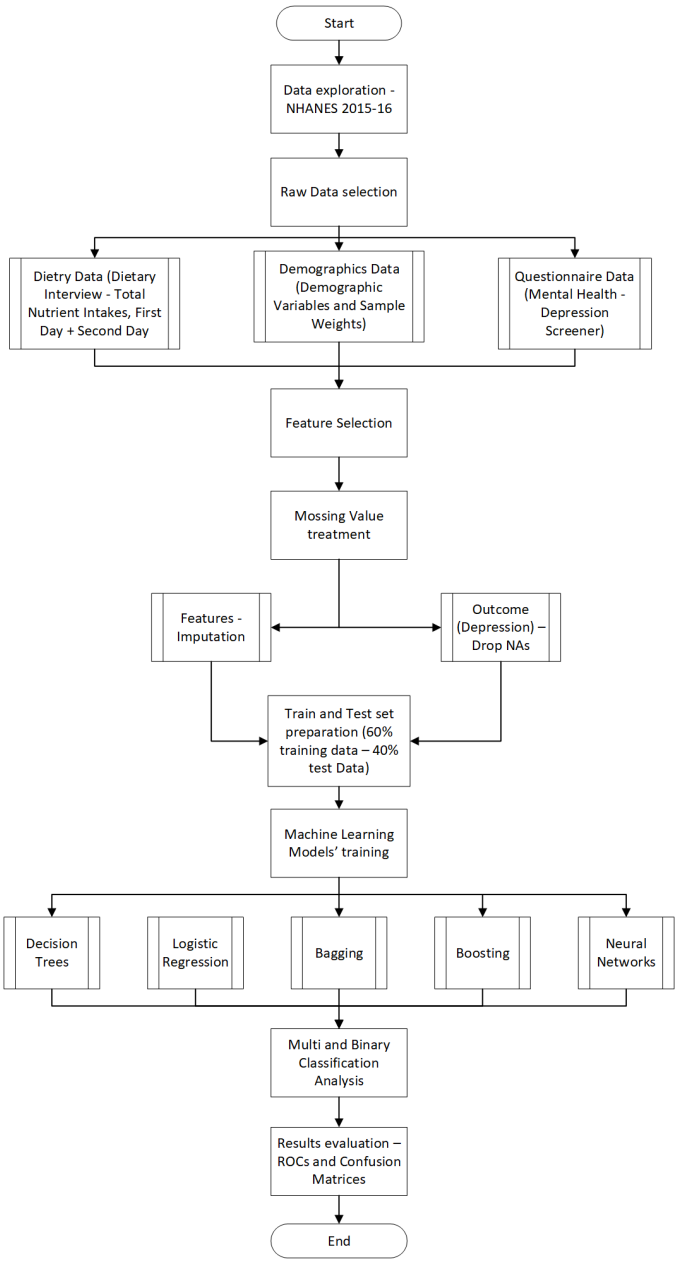



Figure 1: Experimental design flowchart

| Gender | Sample Count |
|---|---|
| Male | 2,759 |
| Female | 2,906 |

| Age | |
|---|---|
| Min | 18 years |
| 1st Quadrant | 32 years |
| Median | 48 years |
| Mean | 48.28 years |
| 3rd Quadrant | 63 years |
| Max | 80 years |

| Ethnicity | Count |
|---|---|
| Mexican | 1,005 |
| Other Hispanic | 740 |
| Non-Hispanic White | 1,822 |
| Non-Hispanic Black | 1,211 |
| Non-Hispanic Asian | 674 |
| Ohter Race | 213 |

| Depression Categorization | Count |
|---|---|
| Minimal Depression | 3,781 |
| Mild Depression | 884 |
| Moderate Depression | 254 |
| Moderately Severe | 101 |
| Severe | 55 |
| Non-Applicable | 590 |

Table 1: Cohort Summary By Gender, Age, Ethnicity and Depression Category based on PHQ-9 Scoring

### 3.2 Data Extraction

Data extraction flow chart is shown below.

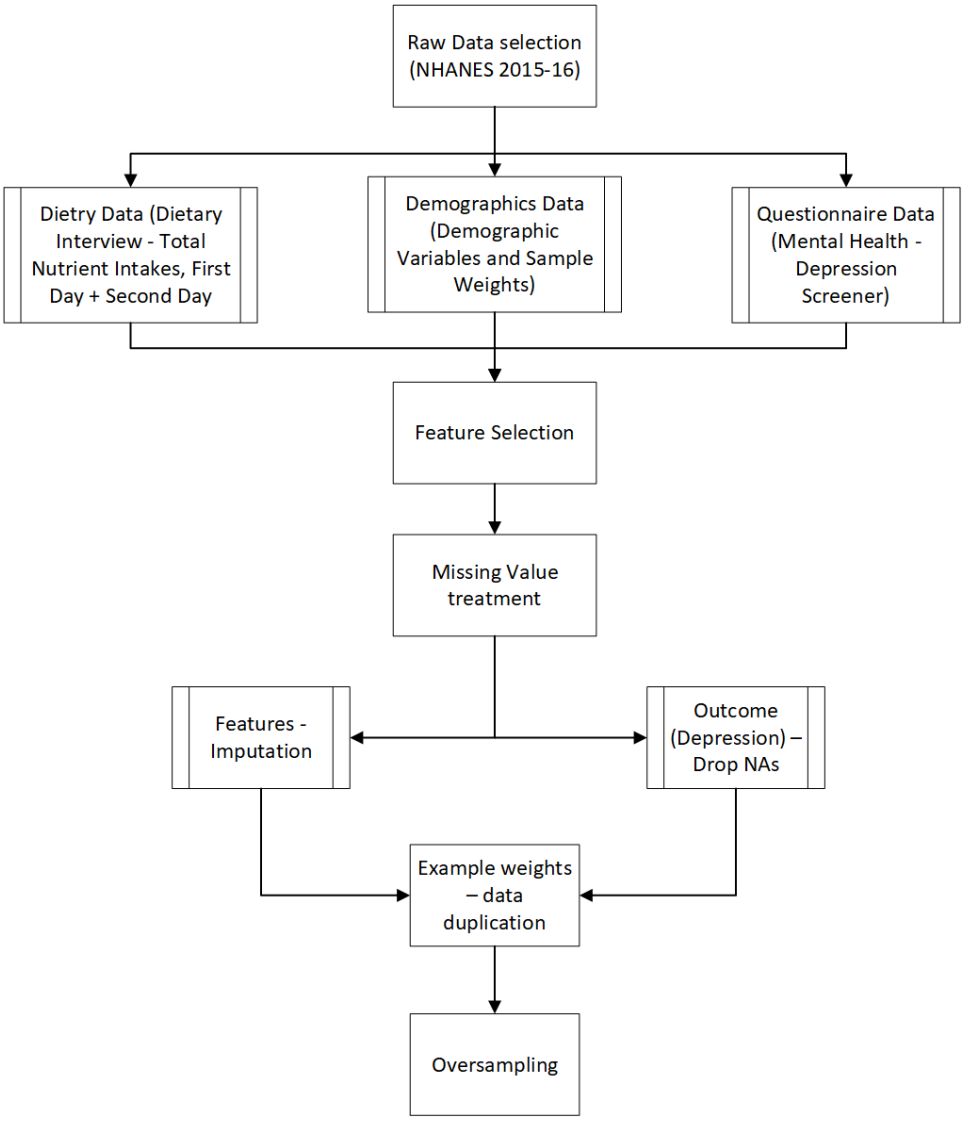



Figure 2: Data extraction

This figure shows our data extraction approach. We used NHANES demographic, dietary, and mental health questionnaire data (PHQ-9). NHANES provides multiple data-sets for dietary data. We used 'Dietary Interview - Total Nutrient Intakes, First Day' (DR1TOT) and 'Dietary Interview - Total Nutrient Intakes, Second Day' (DR2TOT) data-sets. PHQ-9 is a 9-question instrument which is used in primary health-care settings to detect presence and determine degree of depression in patients(Network). The classification of our outcome of interest (depression) is based on the findings of PHQ-9 data provided by NHANES.

Raw data we collected had two major challenges - missing values and unbalanced classes. We dropped all examples where depression data was missing. We imputed missing values for the all the features we had selected. We used oversampling techniques to balance our classes.

NHANES data also had weights for each example which had to be used to ensure our data sample fairly represented the population. We normalized the given weights using a range of 1-10 and duplicated each example accordingly.

### 3.3 Feature Choices

We selected 63 features from all three data-sets. The selected features belong to two categories - demographics and dietary. Demographics provide individual, family, and household-level information. Information such as age, gender, race, education, marital status is included. Dietary data provides detailed dietary intake information about diet components such as sugar, fiber, vitamins, fats as well as information about the type of diet e.g. diabetic, weight loss, gluten free etc.

All nutrition data has been normalized to micrograms. Under demographics, we are only considering adults (18+) because nutrition and depression data is self-reported and our assumption is that adults are able to provide reasonably good estimates of their nutrition. We have also excluded pregnant individuals as their nutrition and mental health is a special case and cannot be generalized for an everyday use-case analysis.

### 3.4 Evaluation Criteria

We have followed two classification approaches - mutli-class classification and binary classification. For some models we only followed one of these approaches, while for certain others we followed both approaches.

We have used confusion matrices for multi-class classification approach. We are interested in each class's TP rate and hence confusion matrix is a suitable option. In confusion matrices we have looked at specificity and sensitivity. We are particularly interested in having sensitivity of moderate to severe depression classes close to 1. We are not considering accuracy as we have highly imbalanced classes and accuracy will be a misleading metric for performance evaluation as we can get a high accuracy by simply predicting everything as the majority class.

We have used receiver operating curve (ROC) for binary classification approach. ROC curves are insensitive to changes in class distribution. Since, we have to oversample data for our analysis, where our class distributions in reality will be imbalanced, therefore, ROC curves are a useful metric for evaluating our binary classification models.

## 4. Methods

(Note: all code described in this section as well as Section 5 is available at `https://github.com/askaricmu/Applied-Analytics-Project-team-1`).

| Depressed | Not Depressed |
|---|---|
| 410 | 4665 |

Table 2: Depression Label Frequency Table

### 4.1 Logistic Regression

The first method we are exploring is binary classification with logistic regression. The participants that were identified with depression levels moderate, moderately severe, and severe were categorized as depressed and those with levels minimal and mild were categorized as not depressed. Table 2 shows the breakdown of participants with and without depression. As shown in table 2 the number of individuals without depression are significantly greater than those with depression. To balance the classes, different oversampling methods were used. For this model, Synthetic Minority Over-Sample Technique (SMOTE) was used to create a more even distribution. We fitted the model to three separate sets of features: (1) SES features; (2) nutrition features; and (3) all features. 5-fold Cross Validations were performed on the models and the confusion matrix and ROC are reported.

### 4.2 Decision Tree

Decision Tree is another model we utilized for binary classification. The participants were identified into two groups and oversampled with the same process that is mentioned above in the logistic regression methodology. As we did with the logistic regression method, decision trees were fitted to three different data sets. We then compared the performance of decision tree model and logistic regression with overlaid ROC curves and accuracy metrics.

### 4.3 Bagging (Bootstrap Aggregation) - Random Forest

We applied Random Forest algorithm to predict various of levels depression. To improve our model, we oversampled the minority classes using ROSE and SMOTE in order to arrive at balanced classes. "ROSE uses smoothed bootstrapping to draw artificial samples from the feature space neighbourhood around the minority class" (MENARDI and TORELLI (a)). "SMOTE draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space" (MENARDI and TORELLI (b)). We analyzed confusion matrix for evaluating our multiclass prediction model.

We converted our multiclass outcome of interest (depression) into a binary class. Moderate, moderately severe and severe classes were encoded as High. Mild and minimal classes were encoded as low. We oversampled this transformed data using both ROSE and SMOTE respectively, and performed predictions through Random Forest in each case. WE analyzed ROC for evaluating binary class prediction model.

### 4.4 Boosting - adaBoost

The modelling steps we followed in Boosting were similar to Bagging except the model itself. We applied adaBoost algorithm to predict various levels of depression. To improve our model, we oversampled the minority classes using ROSE and SMOTE in order to arrive

at balanced classes. We analyzed confusion matrix for evaluating our multiclass prediction model.

We converted our multiclass outcome of interest (depression) into a binary class. Moderate, moderately severe and severe classes were encoded as High. Mild and minimal classes were encoded as low. We oversampled this transformed data using both ROSE and SMOTE respectively, and performed predictions through adaBoost in each case. WE analyzed ROC for evaluating binary class prediction model.

### 4.5 Gradient Boosting

We applied gradient boosting model to predict depression. All other steps of anlysis were similar to adaboost except that we only used SMOTE oversampling in gradient boosting.

### 4.6 K-Nearest Neighbor Modeling

The next method we look into is K-Nearest Neighbor(K-NN) modeling with implementation from the `caret` package in R. We use a 10-fold repeated cross validation on our training set with oversampling of smaller classes, namely those whose depression symptoms are categorized as moderately severe and severe. We first separate the nutritional variables from the SES indicators, then run the K-NN model three times on only nutritional variables, only SES indicators and all variables combined to compare which model performs better. Before fitting our models, we first duplicate the each row according to a normalized sample weight and specify the oversampling.

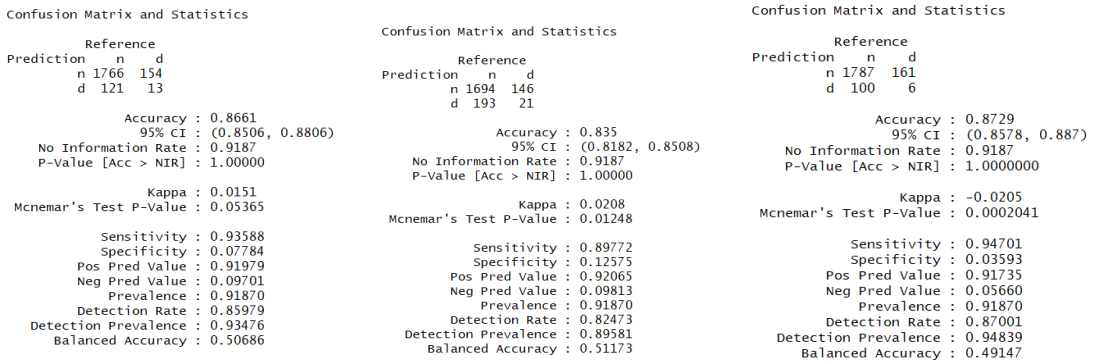
## 5. Results

### 5.1 Results on Logistic Regression

Figure 3 compares the accuracy of logistic regression when modeled on SES, nutrition, and both. All three accuracy metrics (using AUC) rest between 0.83 and 0.87. Comparing these three models, logistic regression on both SES and nutrition performs better than models on SES and nutrition.

| SES Var | Importance | Nutri. Var | Importance | All Var | Importance |
|---|---|---|---|---|---|
| dmdeduc25 | 100.00 | p184 | 100.000 | indfmpir | 100.00 |
| dmdcitzn2 | 74.87 | alco | 88.711 | niac | 63.61 |
| dmdmartl5 | 73.85 | atoa | 34.292 | lz | 60.84 |
| dmdeduc23 | 61.37 | atoc | 29.731 | vc | 60.42 |
| dmdeduc24 | 58.17 | vk | 25.163 | fibe | 54.68 |
| ridreth32 | 56.90 | s100 | 23.814 | fola | 48.08 |
| dmdmartl6 | 55.75 | pota | 23.307 | ret | 46.46 |
| indfmin27 | 55.49 | s140 | 16.710 | iron | 43.14 |
| dmdmartl3 | 48.32 | lyco | 14.900 | m221 | 42.70 |
| indfmin212 | 45.75 | p183 | 14.589 | vara | 40.83 |

Table 3: Logistic Regression Top 10 Variable Importance




(a) Socio-Economic Status (b) Nutrition (c) Total

Figure 3: Confusion Matrices and Statistics of Logistic Regression Models

Table 3 displays the ten predictors that have highest relative importance according to the varImp function in the caret package, which uses the t-statistic for each model parameter.(Kuhn) The result suggests that the 'dmdeduc2' predictor (education level above adults 20+) has the highest variable importance when predicting depression with just SES data. Other variables of high importance are Octadecatetraenoic acid (p814), alcohol consumption (alco), Niacin, also known as Vitamin B3 (niac), potassium (Pota), and vitamin C (VC). Interestingly, comparing the variables in table 3, none of the SES and nutrition variables that were determined as having the highest relative importance are included in the 10 predictors with highest-importance from the model that used both SES and nutrition predictors.

### 5.2 Results on Decision Tree

As shown in figure 4 the accuracy of each model is very similar. Among the three models, decision tree on nutrition data as the highest accuracy of 0.8647, winning by a very small margin. Next, performance of decision tree model and logistic regression is compared in figure 5. Decision tree has a higher AUC when modeled on nutrition data. This can possibly be explained because nutrition has more predictors. Since decision tree partitions the feature space, a non-linear boundary performs better than linear boundary, which is what logistic regression model creates.

```
Confusion Matrix and Statistics

          Reference
Prediction    n    d
         n 1754  146
         d  133   21

               Accuracy : 0.8642
                 95% CI : (0.8486, 0.8787)
    No Information Rate : 0.9187
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0573
 Mcnemar's Test P-Value : 0.4725

            Sensitivity : 0.9295
            Specificity : 0.1257
         Pos Pred Value : 0.9232
         Neg Pred Value : 0.1364
             Prevalence : 0.9187
         Detection Rate : 0.8539
   Detection Prevalence : 0.9250
      Balanced Accuracy : 0.5276
```

(a) Socio-Economic Status

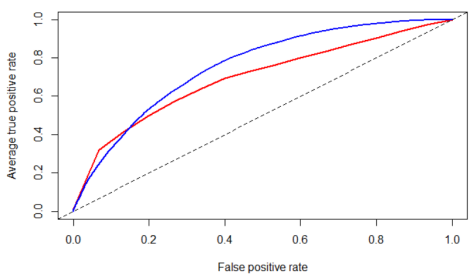
```
Confusion Matrix and Statistics

          Reference
Prediction    n    d
         n 1760  151
         d  127   16

               Accuracy : 0.8647
                 95% CI : (0.8491, 0.8792)
    No Information Rate : 0.9187
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0305
 Mcnemar's Test P-Value : 0.1678

            Sensitivity : 0.93270
            Specificity : 0.09581
         Pos Pred Value : 0.92098
         Neg Pred Value : 0.11189
             Prevalence : 0.91870
         Detection Rate : 0.85686
   Detection Prevalence : 0.93038
      Balanced Accuracy : 0.51425
```

(b) Nutrition

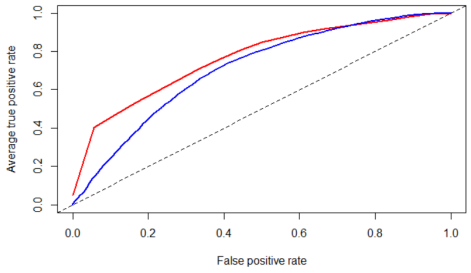
```
Confusion Matrix and Statistics

          Reference
Prediction    n    d
         n 1749  155
         d  138   12

               Accuracy : 0.8574
                 95% CI : (0.8415, 0.8722)
    No Information Rate : 0.9187
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.0013
 Mcnemar's Test P-Value : 0.3499

            Sensitivity : 0.92687
            Specificity : 0.07186
         Pos Pred Value : 0.91859
         Neg Pred Value : 0.08000
             Prevalence : 0.91870
         Detection Rate : 0.85151
   Detection Prevalence : 0.92697
      Balanced Accuracy : 0.49936
```

(c) Total

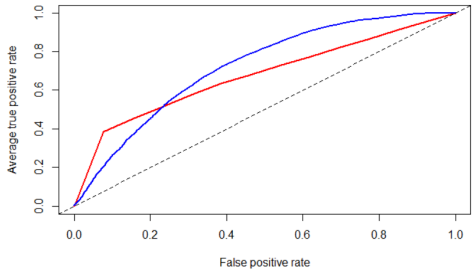
Figure 4: Confusion Matrices and Statistics of Decision Tree Models




(a) Socio-Economic Status




(b) Nutrition




(c) All Features

Figure 5: ROC curves of Logistic Regression (Blue) and Decision Tree (Red)

### 5.3 Results on Bagging (Bootstrap aggregation) - Random Forest

We considered bagging algorithm over Random forest to predict multiclass depression. Since, we had unbalanced classes, we got a worthless but high prediction accuracy - our model predicted everything as majority class. We used confusion matrix to evaluate our model's performance. confusion matrix results are given below.

```
Confusion Matrix and Statistics

                    Reference
Prediction           mild minimal moderate moderately severe severe
  mild                  0       0        0                 0      0
  minimal             509    2339      159                43     37
  moderate              0       0        0                 0      0
  moderately severe     0       0        0                 0      0
  severe                0       0        0                 0      0

Overall Statistics

               Accuracy : 0.7577
                 95% CI : (0.7422, 0.7727)
    No Information Rate : 0.7577
    P-Value [Acc > NIR] : 0.5098
```

Figure 6: Confusion Matrix with Bagging (Random Forest)

As we can see, all examples have been predicted as the majority class (minimal) which gives an accuracy of 75% but is not useful for us.

We tried improving the performance of our model by converting our classes to binary level. We categorized minimal and mild class levels as low, and moderate, moderately severe, and severe class levels as high. In binary classification approach we used ROC to evaluate the performance of our model. We oversampled our data to reduce class imbalance. We used ROSE as well as SMOTE oversampling techniques and compared the results using ROC. We also considered confusion matrix just to get a general understanding of class specific predictions. Confusion matrices and ROCs are given below.

```
          Reference
Prediction High Low
      High    0   0
      Low    76 939

               Accuracy : 0.9251
                 95% CI : (0.9072, 0.9406)
    No Information Rate : 0.9251
    P-Value [Acc > NIR] : 0.5305
```

Figure 7: Confusion Matrix with Bagging - ROSE oversampling

```
          Reference
Prediction High  Low
      High   31  137
      Low   140 1722

               Accuracy : 0.8635
                 95% CI : (0.8478, 0.8782)
    No Information Rate : 0.9158
    P-Value [Acc > NIR] : 1.0000
```

Figure 8: Confusion Matrix with Bagging - SMOTE oversampling

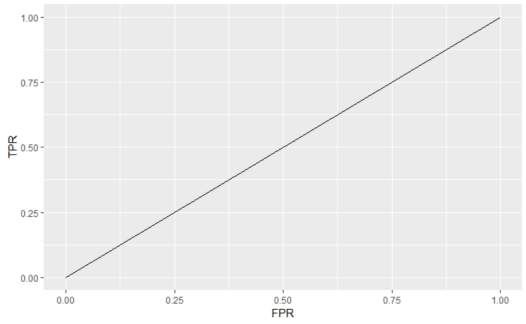



Figure 9: ROC with Bagging - ROSE oversampling

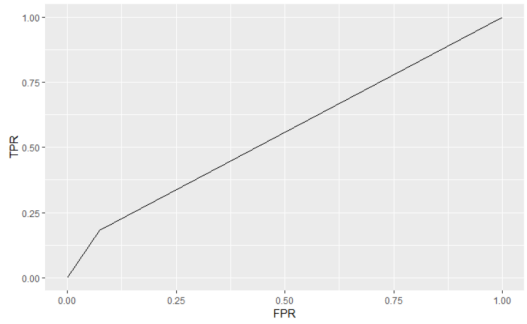



Figure 10: ROC with Bagging - SMOTE oversampling

Looking at these figures, we can observe that binary classification with SMOTE oversampling is atleast able to differentiate between low and high classes. Although the ROCs show that Bagging is not really giving us a usable prediction model, still we can take the learning from this analysis (SMOTE oversampling and binary classification) and build a more sophisticated model for prediction such as Boosting.

### 5.4 Results on Boosting - adaBoost

We used adaboost model from R's ada package to perform binary classification of depression levels. We oversampled the data using SMOTE. Although our bagging model showed that SMOTE worked better than ROSE, we still considered ROSE in this analysis too. Confusion matrices and ROCs for Boosting are given below.

```
          Reference
Prediction High  Low
      High   69  409
      Low   102 1450

               Accuracy : 0.7483
                 95% CI : (0.7288, 0.767)
    No Information Rate : 0.9158
    P-Value [Acc > NIR] : 1
```

Figure 11: Confusion Matrix with adaBoost - ROSE oversampling

```
          Reference
Prediction High  Low
      High   69  456
      Low   102 1403

               Accuracy : 0.7251
                 95% CI : (0.7051, 0.7445)
    No Information Rate : 0.9158
    P-Value [Acc > NIR] : 1
```

Figure 12: Confusion Matrix with adaBoost - SMOTE oversampling

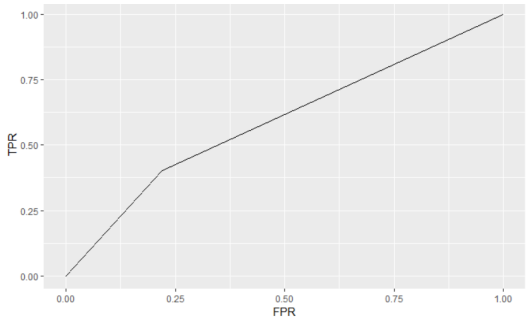



Figure 13: ROC with adaBoost - ROSE oversampling

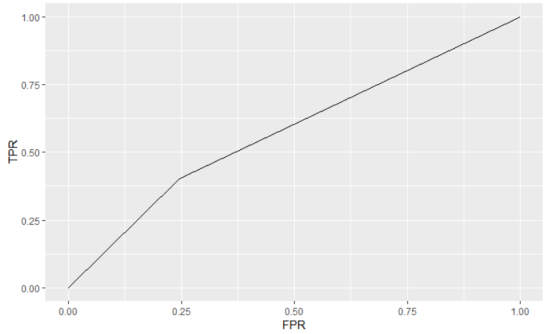



Figure 14: ROC with adaBoost - SMOTE oversampling

Boosting results in similar performance with both SMOTE and ROSE oversampling. However the results in both cases are not good enough for our model to be adopted for prediction in general. Both models have missclassified almost two-thirds of high class. We are interested in predicting high depression with better accuracy. We will not consider gradient boosting to further our analysis.

### 5.5 Results on Gradient Boosting

We used gbm package in R to build a gradient boosting model for binary classification of depression. We used SMOTE oversampling and used ROC to evaluate performance of our model. The ROC for gradient boosted model is given below.

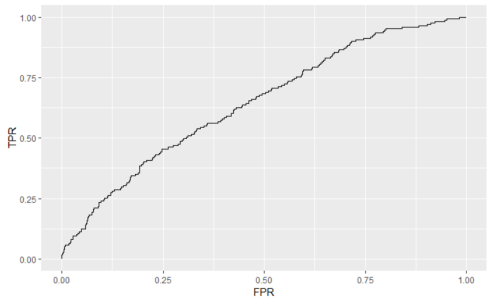



Figure 15: ROC with Gradient Boosting - SMOTE oversampling

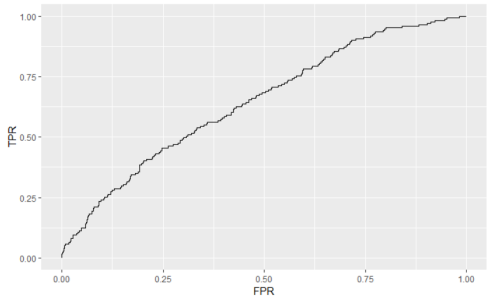



Figure 16: Confusion Matrix with Gradient Boosting - SMOTE oversampling

The ROC produced by gradient boosting appears to be the most reasonable one among all the ROC's produced through our various approaches in bagging and boosting. Nonetheless, this ROC is not very different from the others.

### 5.6 Bagging and Boosting - bringing it all together

We will look at the change in sensitivity across various bagging and boosting approaches that we tried. Figures below show the sensitivity for all our models.

```
                Sensitivity : 0.0000000
                Specificity : 0.9989362
             Pos Pred Value : 0.0000000
             Neg Pred Value : 0.9260355
                 Prevalence : 0.0738916
             Detection Rate : 0.0000000
       Detection Prevalence : 0.0009852
          Balanced Accuracy : 0.4994681
```

Figure 17: Bagging metrics - ROSE oversampling

```
          Sensitivity : 0.0000000
          Specificity : 0.9989362
       Pos Pred Value : 0.0000000
       Neg Pred Value : 0.9260355
           Prevalence : 0.0738916
       Detection Rate : 0.0000000
 Detection Prevalence : 0.0009852
    Balanced Accuracy : 0.4994681
```

Figure 18: Bagging metrics - SMOTE oversampling

```
          Sensitivity : 0.38514
          Specificity : 0.78215
       Pos Pred Value : 0.12206
       Neg Pred Value : 0.94178
           Prevalence : 0.07291
       Detection Rate : 0.02808
 Detection Prevalence : 0.23005
    Balanced Accuracy : 0.58364
```

Figure 19: Boosting metrics - ROSE oversampling

```
          Sensitivity : 0.43243
          Specificity : 0.74814
       Pos Pred Value : 0.11896
       Neg Pred Value : 0.94370
           Prevalence : 0.07291
       Detection Rate : 0.03153
 Detection Prevalence : 0.26502
    Balanced Accuracy : 0.59029
```

Figure 20: Boosting metrics - SMOTE oversampling

```
          Sensitivity : 0.39865
          Specificity : 0.81031
       Pos Pred Value : 0.14183
       Neg Pred Value : 0.94486
           Prevalence : 0.07291
       Detection Rate : 0.02906
 Detection Prevalence : 0.20493
    Balanced Accuracy : 0.60448
```

Figure 21: Gradient Boosting metrics - SMOTE oversampling

Considering the sensitivity, as we are trying to predict high depression cases, adaBoost model with SMOTE oversampling has produced comparatively better results. Bagging is not a suitable model when we have imbalanced classes. We expected boosting and gradient boosting to perform better than they did. However, the dataset we had was not large enough to contain a significant number of minority class examples. We had more than 4000 majority class examples but only a few hundred minority class examples. In the context of our analysis, bagging and boosting may only be used to build a general prediction model if the available data set is large (a few thousand minority class examples).

Top 5 variables in terms of importance in Boosting were s060 (Hexanoic), tfat (Total Fat), atoc (Vitamin E as alpha-tocopherol), fibe (Low fiber diet) and ret (Retinol). This informs us that we should look to further explore these variables to understand their impact on adults suffering from depression. This is just a preliminary finding. However, it is worth exploring further.

### 5.7 Results on K-Nearest Neighbor

Since we are most interested in those with moderate to severe and severe depression, the metrics to determine how effective on the K-NN model is would be the weighted accuracy for those two specific classes, as well as the false negative rate since it is potentially more costly to mistake a depressed individual as depression-free. The results are displayed as the following Confusion Matrices in Figure 22, Figure 23 and Figure 24.

```
                    Reference
Prediction           mild minimal moderate moderately severe severe
  mild                 97     375       27                14      6
  minimal             178     701       39                22      5
  moderate             19     102        5                 5      3
  moderately severe     8      38        3                 1      0
  severe                3       6        0                 1      0

Overall Statistics

               Accuracy : 0.4849
                 95% CI : (0.4606, 0.5093)
    No Information Rate : 0.737
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0054
 Mcnemar's Test P-Value : <2e-16
```

Figure 22: Confusion Matrix with K-NN Model using only Nutritional Data

```
                  Reference
Prediction          mild minimal moderate moderately severe severe
  mild               113     450       43                21      9
  minimal            170     746       32                15      3
  moderate            47     175       12                 7      3
  moderately severe   32      95        6                 2      3
  severe              16      47        3                 2      2

Overall Statistics

               Accuracy : 0.426
                 95% CI : (0.4045, 0.4477)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0257
 Mcnemar's Test P-Value : <2e-16
```

Figure 23: Confusion Matrix with K-NN Model using only SES Data

```
                  Reference
Prediction          mild minimal moderate moderately severe severe
  mild                78     363       22                18      4
  minimal            181     726       35                21      5
  moderate            27      95       11                 2      5
  moderately severe   11      21        4                 1      0
  severe               8      17        2                 1      0

Overall Statistics

               Accuracy : 0.4922
                 95% CI : (0.4678, 0.5165)
    No Information Rate : 0.737
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0069
 Mcnemar's Test P-Value : <2e-16
```

Figure 24: Confusion Matrix with K-NN Model using both Nutritional Data and SES data

None of the three models performs as well as the above ones. There are several possible explanations. First of all, the `caret` package that implements K-Nearest Neighbor uses euclidean distance in high dimensional space to determine the proximity of each data point, and because the data consists of a mixture of categorical variables and numeric variables, the distance between two data points can be hard to quantify. Another issues can be for the fact that the data has too many dimensions. It is hard to train the model without reducing the dimensionalities first.

## 6. Discussion and Related Work

### 6.1 Potential Limitations

One potential limitation of this study is the nature of the data source. NHANES rely on surveys and self-reported mental status.For eaxample, one of NHANES's survey question to screen for depression is: Thinking about the last time you felt worried, nervous or anxious, how would you describe the level of these feelings?. The response is very subjective and this may cause biases and inadequate labeling of a participant as depressed.

Another limitation also arises from NHANES's survey data. The inputs to the analysis is dependent on the NHANES questionnaire. Therefore, we are limited to the information that is provided by NHANES where there may be other underlying SES factors. Additionally, the most recent set of survey results are from 2016. As there is a time delay of 2 years, the analysis may not be reflective of current ground truth.

### 6.2 Neural Network Modeling Approach

In addition to models presented in previous sections of this paper, we also explore the possibility of training a neural network. Inspired by interconnected neurons in biological systems, neural networks are powerful frameworks that allows multiple machine learning algorithms to process complex data units (n.d. (2017)). For this research purpose, we intend to use a multi-layered network to train our data on. We use softmax for our output encoding for our k-ary classification and use multi-class cross entropy as our objective function for the encoding.

However, there are a few issues that prevent us from utilizing a 'Keras' sequential model in R for this project effectively. Neural Networks usually out-perform traditional models when there is a huge amount of data. With only about 5,000 entries of data, we find the performance of a neural network unable to surpass other models. Secondly, like K-NN model, neural networks are unable to analyze a data set with both numeric data and categorical data. `Keras` also does not come with parameters for us to address the issue of data imbalance. Therefore, we welcome future endeavors to use neural networks for identifying the relationships between SES, nutrition and Depression. But for this paper we would like to withhold our findings using a neural network.

### 6.3 Other Related Works

In addition to the works mentioned in Section 1, there is an overabundance of literature either directly or tangentially related to our research. A 2017 article by Shervin Assari titled "Social Determinants of Depression: The Intersections of Race, Gender, and Socioeconomic

Status" has examined the relationship between SES and major depressive episodes within a 12-month period using a representative sample of the U.S adult population (Assari (2017)). For links between nutrition and depression, a report using 2007 - 2012 NHANES data by M. Soledad Cepeda, MD, PhD and co. claims that depression is associated with high levels of C-reactive protein and low levels of fractional exhaled nitric oxide (M. Soledad Cepeda et al. (2016)). Those works are all very insightful and helpful in guiding our research when it comes to selecting cohorts and building models.

## 7. Conclusion

Multiple modeling techniques were explored in the study, ranging from binary classification with logistic regression, decision tree, ensembling, and K-NN. There were no clear winning models based on our results. Logistic regression, decision tree, ensembling models (ada Boost, bagging) all yielded similar results and accuracy. However, simpler models performed slightly better than the more complex models. The performance of the models can potentially be improved by using data sets that are better balanced and more stable. For example, instead of using a self-reported evaluation survey, clinical data could provide more robust results. Despite not having developed a definite model that can classify depression from SES and nutrition with high precision, there is clear evidence of the effects of SES and nutrition on an individuals mental health. Therefore, studies should be continued to further understand these impacts on depression by exploring new data sets and algorithms.

## References

*Modern Applied Statistics with S.*

N.S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46:175–185, 1992.

Shervin Assari. Social determinants of depression: The intersections of race, gender, and socioeconomic status. *Brain Sciences*, 7, 2017.

J. Brownlee. Bagging and random forest ensemble algorithms for machine learning. 2016.

N Darmon and A Drewnowski. Does social class predict diet quality? *The American Journal of Clinical Nutrition*, 87:1107–1117, 2008.

S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Biomedical Informatics*, 35:352–359, 2002.

J. D'Souza. A quick guide to boosting in ml. 2018.

R Ahrnsbrak et al. Key substance use and mental health indicators in the united states: Results from the 2016 national survey on drug use and health. 2017.

M.A. H Hedegaard, M.D.and S Curtin and Ph.D. M Warner. Suicide rates in the united states continue to increase. 2018.

Max Kuhn. *Classification and Regression Training (caret).*

PhDa M. Soledad Cepeda, MD, Paul Stang PhD, and MSa Rupa Makadia. Depression is associated with high levels of c-reactive protein and low levels of fractional exhaled nitric oxide:results from the 20072012 national health and nutrition examination surveys. *The Journal of Clinical Psychiatry*, 77, 2016.

G. MENARDI and N. TORELLI. Training and assessing classification rules with unbalanced data. *Working Paper Series*, a.

G. MENARDI and N. TORELLI. Training and assessing classification rules with unbalanced data. *Working Paper Series*, b.

K Miyaki, Y Song, and S Taneichi et al. Socioeconomic status is significantly associated with the dietary intakes of folate and depression scales in japanese workers (j-hope study). *Nutrients*, 5:565–578, 2013.

n.d. Build with ai. 2017.

Psych Congress Network. Patient health questionnaire (phq-9).

R Pechey and P Monsivais. Socioeconomic inequalities in the healthiness of food choices: Exploring the contributions of food expenditures. *Preventive Medicine*, 88:203–209, 2016.