Final Project Report

National Health and Nutrition Examination Survey

Toby Anderson

09/17/2021

# Introduction

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.[[1]](#footnote-1) The survey is given each year to about 5,000 participants from a wide distribution of counties across the United States. It is designed to include demographic, socioeconomic, and dietary data alongside disease indicators. This design allows researchers to find causal links between disease trends and risk factors.

Over 50,000 medical research papers have cited the NHANES data.[[2]](#footnote-2) The data is commonly used by researchers to determine factors contributing to disease risk. Additionally, the data is used to examine dietary patterns, study comorbidities, and find disease trends. By finding relationships between diseases and risk factors, public health advice can improve. And with this open data, the public can verify the relationships and modify their lifestyles to reduce disease risk and improve their expected lifetime and quality of life.

# Analysis and Models

## About the Data

The NHANES survey given in 2013-2014 includes over 1800 data points for each participant, and nearly 10,000 participants. The data is coded, for instance, some example column names include: BPXSY1, CBD090, DBD905. The prefixes to the column group the questions into sections about different topics, for instance, DIQ prefixed columns are about diabetes risk, DMD prefixes are demographic data and DR prefixes are dietary data. In this analysis, many of the columns were translated into more human-readable names, such as: dietary\_b12, low\_carb\_diet, poverty\_level\_index.

The values in each column are all numeric, however, some of the columns represent category data, while others represent ordinal/continuous data. Both the categorical and continuous data contain dummy values to allow for participants who either refused to answer or didn’t know, see Figure 1 below. The dummy values are intentionally extreme outliers and introduce significant bias into the data. These were processed by treating them as Null values and either removed or replaced with the mode of the column.

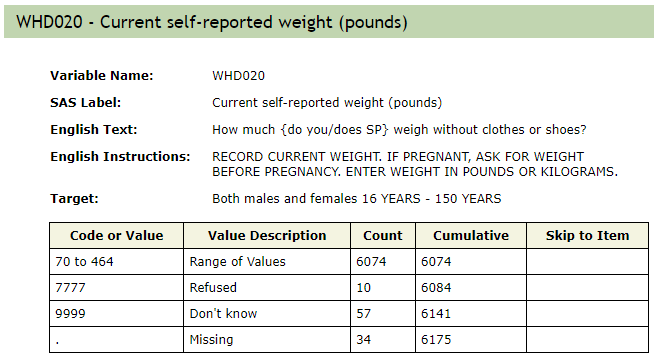


Figure 1 – Example of question with dummy values

The categorical data required further cleaning. The values in these columns cannot be assumed as ordinal, with a relationship that increases with the designated value, see Figure 2. Instead, these were turned into binary columns, with the most common category coded as a “1” and all other options coded as a “0”. The ordinal columns had a similar problem, with dummy values for selections such as “other”, so these outliers were treated as Nulls.

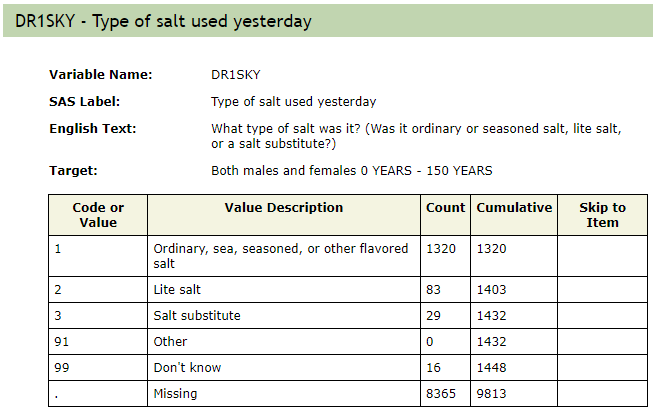


Figure 2 – Example of categorical data that cannot be interpreted as ordinal

After removing outliers and dummy values, Nulls were replaced with the mode of the column. With these replacements, the columns fit an expected distribution.

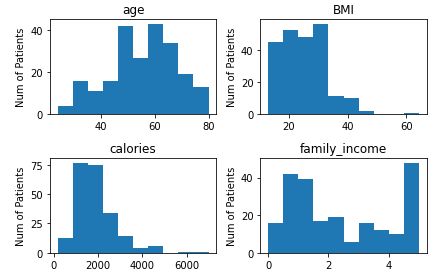


Figure 3 – Histograms of explanatory variables. Note that family\_income is an ordinal column

While many diseases were indicated in the survey, this analysis is limited to four: Liver Conditions, Diabetes, Cancer, and Coronary Heart Disease. All types of liver conditions were lumped together as were all types of cancer. The data was cleaned, filtered, and transformed separately for each target disease in order to reduce the chance of data leakage.

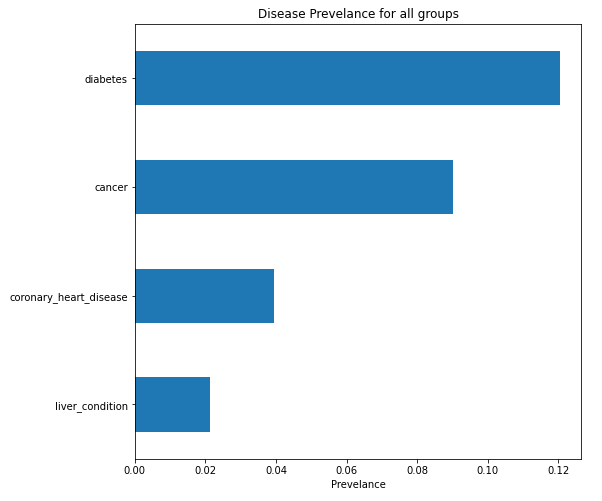


Figure 4 – Prevalence of diseases among survey participants

Although there was much colinearity among the explanatory variables, this was not cleaned before modeling. Instead, regularization was used within models whenever possible in order to reduce the number of redundant variables and reduce the risk of inducing a reversed correlation. Much of the correlation between explanatory variables can be seen within the dietary factors. One explanation is that different types of food come with their own profile of micronutrients. For instance, if someone was to eat many leafy green vegetables, they would consume a high amount of vitamin K, folate, and beta carotene. Since micronutrients come packaged together in macro foods, they will inevitably show correlation, see figure 5.

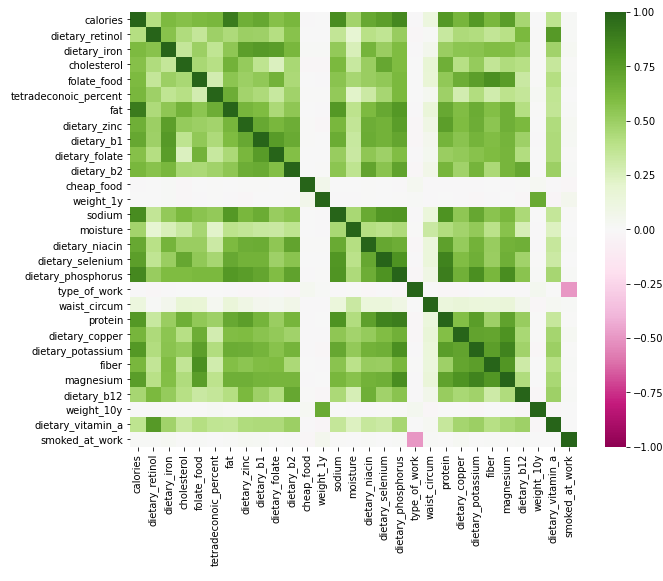
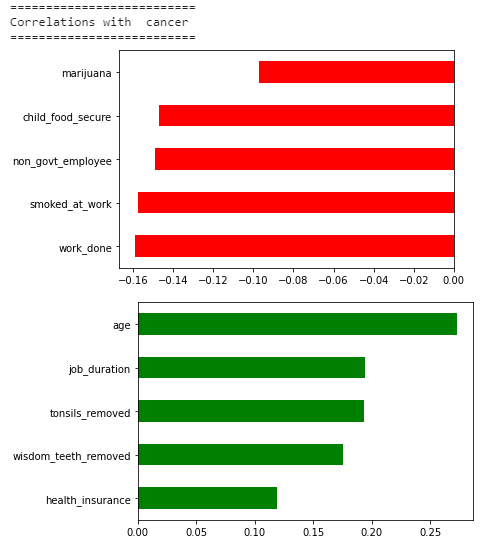
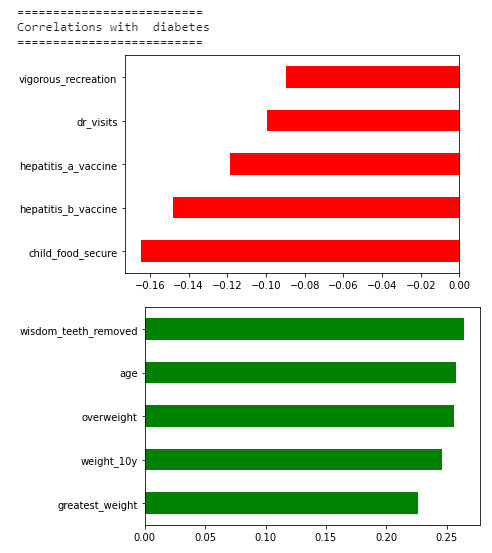


Figure 5 – Strongest correlations between explanatory variables, mainly seen between dietary factors

Finally, one of the main issues with the data was target leakage. Target leakage is the use of an explanatory variable in a model which would not be available at the time of prediction. For instance, one of the factors in this model was the use of insulin. The use of insulin is highly correlated with diabetes risk. However, it is not useful as a predictive factor, because insulin is used to treat diabetes, so if used, the patient would already know they had diabetes. In this way, either treatments for disease, or methods for detecting diseases (such as blood levels of glucose) had to be cleaned out of the data to reduce target leakage. The method for this cleaning was to manually select lifestyle factors to include in the data, which would not indicate pre-knowledge of a disease diagnosis. Correlations between the selected lifestyle factors and diseases are shown below.



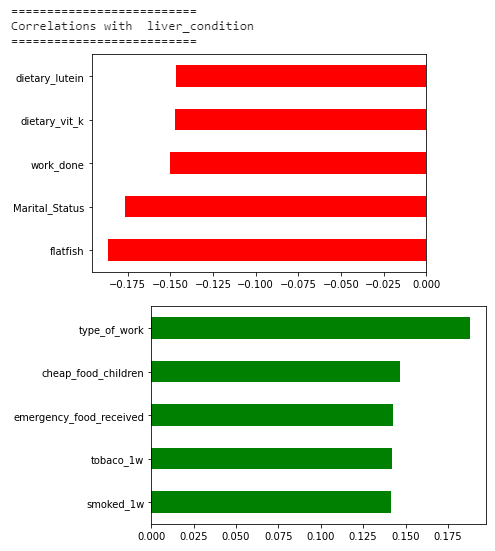
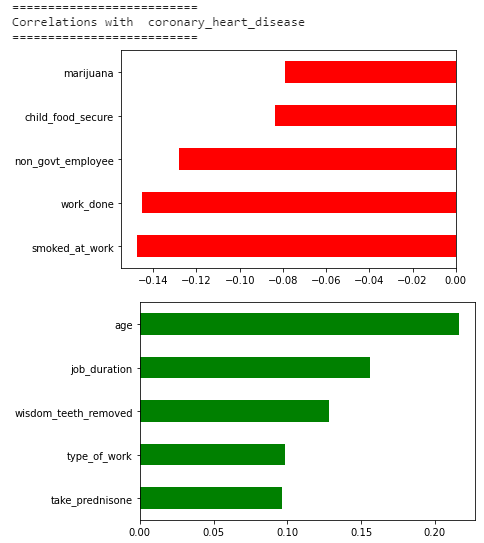


Figure 6 – Correlations between lifestyle factors and diseases

## Models

### Decision Tree

Decision Trees were used to find the most important factors needed for predicting the four target diseases. Although Decision Trees may not be the most sophisticated models for provided accurate prediction results, they provide a convenient way to visualize feature importance. In each Decision Tree, a class weight was provided to balance the data, and a cost complexity pruning factor of 0.0001 was used. The visualizations below were given a max depth of 3. The Decision Tree performed well on Liver Disease and Diabetes, so these results are given below.

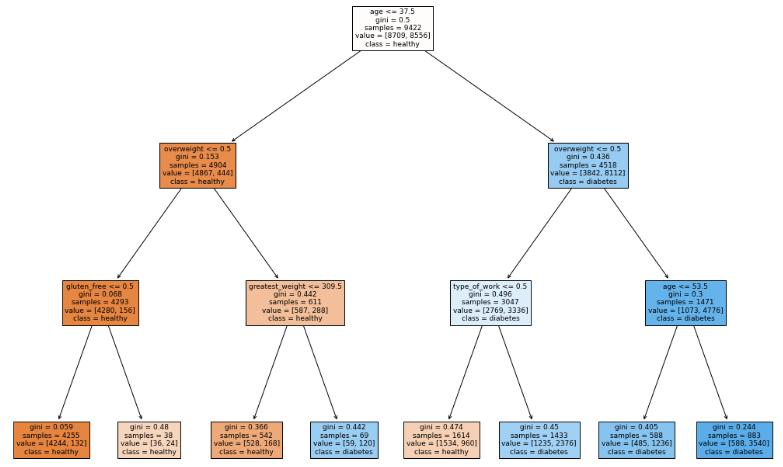


Figure 7 – Diabetes Decision Tree

The Diabetes Decision Tree shows that age and body weight are the most significant factors in determining disease risk. Working also reduces diabetes risk, however, a gluten free diet may increase risk.

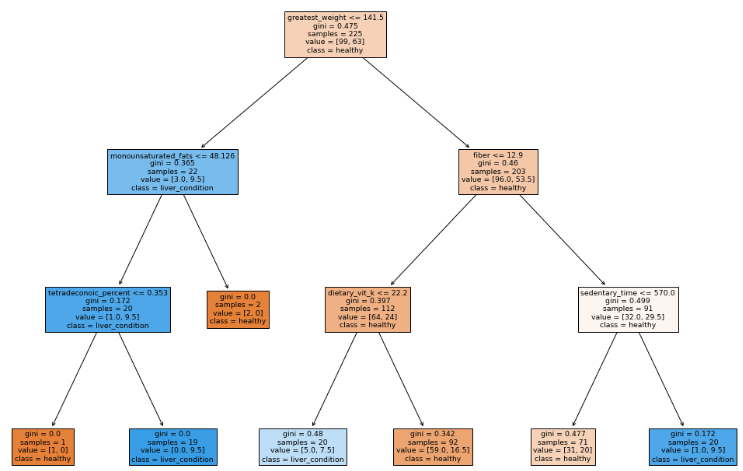


Figure 10 – Liver Disease Decision Tree

In the Liver Disease Decision Tree, age is not the primary factor, which is a surprising result. Low body weight will decrease risk of liver disease. After weight, low fiber with high dietary vitamin K, or high fiber with physical activity decreases cancer risk. For low weight individuals, having high monounsaturated fats, or low saturated fats protects against liver disease.

### Naïve Bayes

A Gaussian Naïve Bayes model was used on all four target diseases with confusion matrices shown below. A smoothing value of 1e-7 was used. The NB models performed poorly except when predicting Cancer. The variable importances for Cancer predictions are shown below.

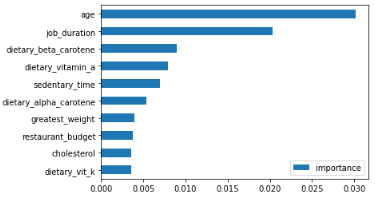


Figure 11 – Variable importances for NB prediction of Cancer

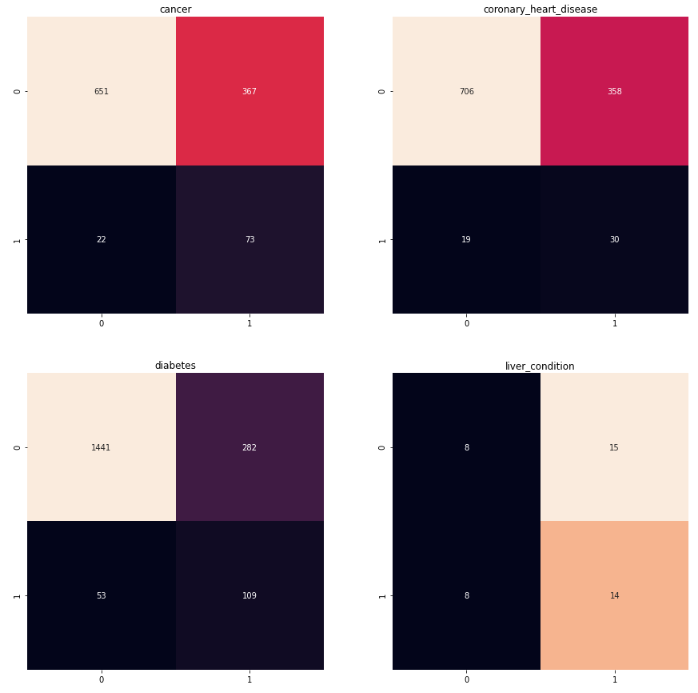


Figure 12 – Confusion Matrices for NB disease prediction

The variable importances show that age is the most influential factor in determining cancer risk. Job duration is also highly correlated with age. Other factors include Vitamin A and Beta Carotene consumption. Perhaps counter-intuitively, higher vitamin A consumption increase risk of cancer. One possible explanation is that most vitamin A consumption is from animal products, and processed meats could be linked to colon cancer.

### Support Vector Machine

To determine the best kernel, cost, and degree for a Support Vector Machine model, an automatic hyper-parameter tuning method was used, which optimized the F1 score. However, even with parameter optimization, the Support Vector Machine performed very poorly on target diseases. The resulting scores are given in the Results section, however, the important features are not included in this report, as the model performance was low. For completeness, the confusion matrices are included below.

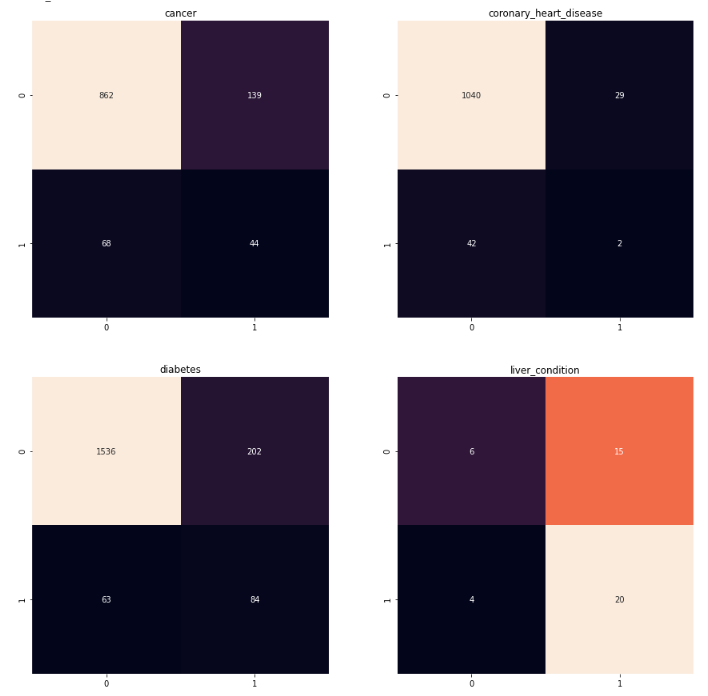


Figure 13 – Confusion Matrices for Support Vector Machine predicitons

### XGBoost

In order to predict Coronary Heart Disease effectively, an XGBoost model was used, which is a gradient boosting model. Gradient boosting is a type of ensemble method, where new models are added in order to predict the errors from previous models. The ROC for coronary heart diseases is provided below.

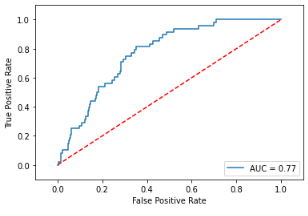


Figure 14 – XGBoost ROC curve for predicting coronary heart disease

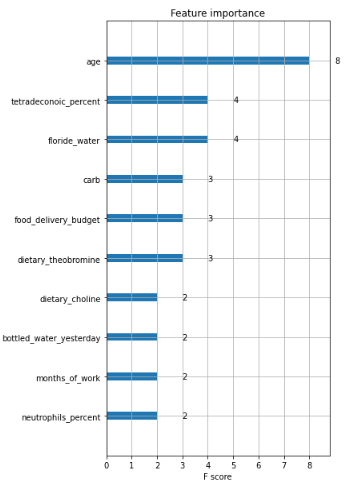


Figure 15 – Feature importance for XGBoost model

The feature importance shows that age is the most important factor. Additionally, dietary saturated fat, floridated water, carbs, and delivered food are important factors when predicting heart disease.

# Results

The F1 scores for all models are given in Table 1. Since the target data is assymetrical, F1 is one of the harshest ways to measure the performance of the models. Overall, Liver Disease was the easiest to predict, followed by diabetes. Coronary heart disease was the most difficult disease to predict. The correlation between age and body weight found in the data held up when examining the important factors of each model. However, factors such as the removal of the tonsils and wisdom teeth, which showed high correlations with disease, were dropped out in favor of more important factors within the models. This may indicate that there is still a problem of collinearity among the explanatory variables. If reproducing this analysis in the future, it would be worthwhile to run models against smaller subsets of factors and be more careful when choosing which features to include.

|  |  |  |
| --- | --- | --- |
| Model | Target | F1 Score |
| Decision Tree | Diabetes | 0.33 |
| Cancer | 0.26 |
| Heart Disease | 0.14 |
| Liver Disease | 0.69 |
| Gaussian NB | Diabetes | 0.28 |
| Cancer | 0.32 |
| Heart Disease | 0.14 |
| Liver Disease | 0.67 |
| SVM | Diabetes | 0.06 |
| Cancer | 0.07 |
| Heart Disease | 0.01 |
| Liver Disease | 0.71 |
| XGBoost | Diabetes | 0.42 |
| Cancer | 0.29 |
| Heart Disease | 0.19 |
| Liver Disease | 0.70 |

Table 1 – Summary of results

# Conclusions

A common theme among all models was the significance of age in disease risk. This is not necessarily a surprising conclusion, but it does lend support to the validity of the data used in this report. Unexpected results included:

* Sleep quality and duration were not major factors in disease risk
* Intense Exercise was not a major factor
  + However, working in general or walking to work reduced disease risk
* Dietary strategies such as low carb, high fiber, or high protein were not factors
  + Quality and type of food were major factors, including cooking at home and consuming high amounts of nutrients found in vegetables
* Smoking had mixed results
  + It appears that smoking has a moderating effect on body weight, so may cancel out it’s overall influence on disease risk.

Overall, having low body weight, staying moderately active, and eating vegetables will reduce risk of diabetes, cancer, heart disease, and liver disease.

1. https://www.cdc.gov/nchs/nhanes/about\_nhanes.htm [↑](#footnote-ref-1)
2. https://pubmed.ncbi.nlm.nih.gov/?orig\_db=PubMed&term=NHANES&cmd=search [↑](#footnote-ref-2)