Determinants of Perceived Health: Dietary Habits, Socioeconomic Factors, Physical Activity, and Personal Metrics

A 2019 and 2021 SMART BRFSS City and County Data Analysis

Author: Obinna Dinneya

Date of Submission: November 27th, 2023

Supervisor: Tamer Abdou

Table of Contents

[Abstract (Modified) 3](#_Toc151742200)

[Intoduction 4](#_Toc151742201)

[Research Questions 5](#_Toc151742205)

[Methodology and Study Design: 10](#_Toc151742209)

[Statistical Analyses: 15](#_Toc151742210)

[Findings and Interpretation: 17](#_Toc151742211)

[Limitations: 27](#_Toc151742212)

[Conclusion: 28](#_Toc151742213)

[References 30](#_Toc151742214)

[Github Links: 31](#_Toc151742215)

## Abstract (Modified)

In today's evolving healthcare landscape, understanding the myriad determinants of perceived health becomes indispensable. Perceived health, while a subjective assessment of one's health status, often mirrors tangible health outcomes, emphasizing its significance in broader healthcare discussions. Historically, researchers have evaluated these determinants in isolation; however, their interconnected nature calls for an integrative approach.

Our research centers around critical research inquiries:

How do dietary habits, specifically the consumption of fruits and green vegetables, relate to overall health outcomes and perceived health predictions?

How does socioeconomic status, emphasizing racial and ethnic backgrounds, influence healthcare access and overall health outcomes?

How does regular physical exercise influence tangible health outcomes, such as the risk of developing cancer, and mental health outcomes, including the likelihood of being diagnosed with depression?

Is there a correlation between chronic conditions and individuals' self-assessed health status?

Can perceived health be accurately predicted using dietary habits, socioeconomic indicators, lifestyle choices, and individual metrics?

To explore these inquiries, we utilize data from the 2019 and 2021 SMART: BRFSS City and County datasets, with 152 variables for 2021 and 161 for 2020(Centers for Disease Control and Prevention, 2021). Our methodology is two-pronged: first, emphasizing the complete set of shared original features from both datasets for predictive modelling and, subsequently, contrasting the models' performances with the full feature set against a select subset (based on importance in predicting perceived health).

By amalgamating these determinants, from dietary choices to socioeconomic contexts, our study endeavours to paint a comprehensive picture of the factors driving perceived health. The anticipated findings aim to direct personalized healthcare strategies, steer policy alterations, and identify potential areas ripe for future research, all with the ultimate aspiration of promoting the holistic well-being of diverse populations.

## Intoduction

## Navigating the vast world of health data today feels like piecing together a complex puzzle. One aspect of health which often gets overlooked is self-perceived health. Why do some individuals feel healthier despite medical ailments while others feel otherwise, even in prime physical conditions? It is like comparing two fruits from the same basket; they might look the same, but their tastes could differ based on various invisible parameters. It is not just about how often we hit the gym or whether we have had our veggies; it is about how all these factors shape our overall sense of well-being.

## Our study dives deep into this, unpacking the little intricacies that paint the bigger picture of perceived health. We are not just talking about the prominent parts, like diet or exercise, but also the subtle factors, like where we come from and the day-to-day challenges we face.

Factors such as our origin, cultural backgrounds, socio-economic status, and the daily hurdles we navigate play an equally pivotal role in shaping our sense of well-being.

## By weaving together these threads—from the food on our plates to the societies we live in, from our workout routines to our battles with chronic conditions—the hope is to present a tapestry that depicts the rich landscape of perceived health. Now, with our roadmap laid out, let us answer the research questions and, more importantly, seek to understand the nuances of perceived health.

## Research Questions

**How do dietary habits, specifically the consumption of fruits and green vegetables, relate to overall health outcomes and perceived health predictions?**

Universally recognized as vital components of a healthful diet, fruits, and vegetables (F&V) are prominent. As underscored by the 2015-2020 U.S. Dietary Guidelines for Americans, it is recommended that F&V constitute one-half of the plate at each meal. This diverse collection of plant foods provides varying energy levels, nutrients, and dietary bioactives essential for human health (Wallace et al., 2020).

Beyond merely meeting our basic nutritional requirements, F&V has shown potential health-promoting effects. They play roles in reducing inflammation and aid in the prevention of various chronic disease states. Fruits and vegetables reduce years lost due to premature mortality and morbidity. With current global intakes of F&V being below recommended levels, there is a pressing need for public policies promoting dietary interventions to help increase F&V intake (Wallace et al., 2020).

In their comprehensive narrative umbrella review, Wallace et al. (2020) delved deep into the clinical and observational evidence on current intakes of F&V. They discussed the available evidence regarding the health benefits of F&V, emphasizing the significant role F&V plays beyond just fulfilling basic nutrient requirements in humans.

A critical takeaway from their research focuses on cardiovascular diseases (CVDs). The review suggests that among the myriad health benefits of F&V, they exhibit the most potent preventive effects against CVDs, mainly when consumed in quantities around 800 grams per day—roughly five servings. It is also noteworthy that certain types of F&V, such as cruciferous vegetables, dark-green leafy vegetables, citrus fruits, and dark-coloured berries, showcase superior effects on biomarkers and outcomes of chronic disease (Wallace et al., 2020). This project aims to uncover whether there is a significant connection between F&V consumption and both objective health outcomes and subjective perceived health.

## How does socioeconomic status, emphasizing racial and ethnic backgrounds, influence healthcare access and overall health outcomes?

In the U.S., significant disparities in healthcare access exist, especially related to socioeconomic and racial factors. By analyzing national survey data from 2011-2015, Towne Jr. (2017) reported that racial and ethnic minority working-age adults, specifically Hispanic adults, were more likely to forgo necessary medical care due to costs when compared to their White counterparts. This discovery was consistent even after adjusting for various other factors, including income, education, and region. Furthermore, individuals with lower incomes or without a college or technical school degree were likelier to skip medical care. Regionally, those residing in the southern U.S. faced higher instances of forgone medical care. State decisions regarding Medicaid Expansion also played a role; individuals in states that did not expand Medicaid reported higher instances of forgone care. Notably, among older adults (65 and above), racial or ethnic minority groups were more likely to forgo medical care than White older adults, highlighting the persistent racial disparities in healthcare access across different age groups. A goal of this project is to investigate the impact of socioeconomic factors, particularly those related to race and ethnicity, on healthcare access and the resulting effects on prediction of overall perceived health.

## How does regular physical exercise influence tangible health outcomes, such as the risk of developing cancer, and mental health outcomes, including the likelihood of being diagnosed with depression?

Regular physical exercise is important in promotion of one’s overall health. It goes beyond merely enhancing physical well-being and playing a pivotal role in mental health. A deeper dive into the literature provides illuminating insights.

Depression and anxiety, prevalent psychiatric conditions, afflict millions in the United States. While there are numerous treatments available, not all are uniformly effective. Intriguingly, a study by Carek, Laibstain, and Carek (2011) postulates that physical activity strongly correlates with decreased symptoms of these conditions. Consistent exercise improves physical health, life satisfaction, cognitive function, and psychological well-being and acts as a buffer against psychological disorders. This research compellingly suggests that exercise is comparable to antidepressant medications for mild to moderate depression and can enhance the effectiveness of such medications. Moreover, while it has been less studied, exercise emerged as a cost-efficient treatment for various anxiety disorders.

Expanding this perspective, Pedersen and Saltin (2015) examined how exercise is a therapeutic tool for a gamut of 26 chronic diseases, including psychiatric, neurological, metabolic, cardiovascular, pulmonary, musculoskeletal disorders, and even cancer. Their analysis gleaned from many sources, including systematic reviews, meta-analyses, and randomized controlled trials, heralds exercise as a pivotal intervention. In many instances, exercise therapy is on par with, if not superior to, medical treatments in efficacy.

The power of regular physical exercise on tangible health outcomes and mental well-being is quite evident. As our understanding of its multifaceted benefits deepens, it underscores the need for integrating exercise into therapeutic regimens and broader public health initiatives. The exploration seeks to unveil the multifaceted effects of regular physical activity on both physical health, particularly the risk of cancer, and mental health, including its role in mitigating the likelihood of depression.

## Is there a correlation between Chronic conditions and individuals self-assessed health status?

When it comes to understanding how we see our health, it's more than just a feeling. It's fascinating how our self-view of our health can tell us about future doctor visits or even how our longevity might come into play.(Palladino et al., 2016).

Various studies present an investigation of the intricate relationship between chronic diseases and self-assessed health status. Barreto & Figueiredo (2009) delved deep into this association, focusing on how gender can influence these perceptions. Their research encompassed 39,821 adults, revealing a noteworthy association between the number of chronic diseases and self-perceived health. They discerned that individuals with a higher number of risk behaviours had a reduced reporting of two or more chronic diseases, suggesting the possibility of reverse causality or enhanced survival rates in those who practice better self-care. This result begs the question: Do folks start living healthier after a diagnosis? Or do those who care for themselves have better chances of beating the odds of chronic illnesses?

The OECD report further sheds light on the subjective nature of health assessments across nations. Most OECD countries have a majority of adults reporting good health. However, countries such as Japan, Korea, Latvia, and Portugal show a significant proportion of adults assessing their health as subpar. Several factors contribute to this variation, including socio-economic conditions, risk factors like smoking, and even financial barriers to healthcare access. Socio-economic disparities, in particular, create a pronounced gap; people with higher incomes consistently reported better health than their lower-income counterparts. This finding reinforces the potential bidirectional relationship: good health leads to better income opportunities, while better income affords better healthcare and lifestyle choices.

One crucial takeaway from the OECD report is caution when comparing perceived health statuses internationally. Cultural, socio-economic, and even survey methodologies can significantly influence the respondents' perception, making direct comparisons challenging. Yet, the recurring theme remains consistent: chronic conditions and individuals' perceptions of their health status are closely entwined, influenced by many factors ranging from individual behaviours to societal structures. Using BRFSS Data this inquiry aims to establish whether individuals with chronic conditions tend to perceive their health differently and, if so, to what extent.

**Can perceived health be accurately predicted using dietary habits, socioeconomic indicators, lifestyle choices, and individual metrics?**

The way individuals perceive their health status is a multi-faceted concept. A particularly insightful study on this topic comes from Teresia Mbogori and Tya M. Arthur, titled "Perception of Body Weight Status Is Associated with the Health and Food Intake Behaviors of Adolescents in the United States."

In this cross-sectional research, the aim was to explore the relationships between adolescents' perception of their body weight status, their self-reported health status, the quality of their diet, and their consumption patterns of fruits and vegetables. The findings from 1737 adolescents aged 12-17 who participated in the Family Life, Activity, Sun, Health, and Eating study. 62% of the participants felt their weight was "just right." Contrastingly, 10.9% considered themselves "underweight," while 22.4% and 4.7% perceived themselves as "a little overweight" and "very overweight," respectively.

An essential takeaway from the study was the positive correlation between weight perception and diet quality. Adolescents who believed their weight was "just right" were more likely to describe their health status as either "very good" or "excellent." This group also reported having a good-quality diet. A key observation was the dietary habits related to fruit and vegetable intake. Those perceiving their weight as "just right" consumed fruits and vegetables more frequently than their peers who saw themselves as either "underweight" or "overweight."

This research by Mbogori and Arthur underscores the crucial relationship between self-perception of body weight and adolescent dietary behaviours. It provides a lens through which we can appreciate the nuanced interplay between perception, health status, and dietary habits. It also stresses the importance of considering self-perception when discussing diet quality and overall health.

The prediction of perceived health indeed intertwines with multiple factors. As highlighted by the study, one's weight perception can influence and, in turn, be influenced by their dietary choices, race and gender. The endeavor is to ascertain the feasibility of predicting perceived health based on a comprehensive set of factors, encompassing dietary habits, socioeconomic status, lifestyle choices, and individual characteristics.

## Methodology and Study Design:

The analysis incorporated data from the Behavioral Risk Factor Surveillance System (BRFSS), specifically using the public use files for the years 2019 and 2021, which included 210,771 and 227,191 non-institutionalized adult respondents, respectively (Centers for Disease Control and Prevention [CDC], 2021). This annual telephone survey collects a wide array of health-related information from adults living in the United States. For this project, 2019 and 2021 Data were converted from XPT files to CSV files using R haven package. Both datasets were read into python and underwent preliminary cleaning and preprocessing using Python libraries (pandas and numpy). The preprocessing techniques included obtaining common columns, renaming, data imputation and removal of all missing values. The 2019 Income, Healthplan, Medical cost, Hypertension columns were renamed to match the 2021 columns due to updates made to variable codenames. Pandas was used to employ stratified mean ,median or mode imputation techniques to handle missing values of important columns. Finally, all rows with missing values were removed. This resulted in two datasets; 2019 (n=83,666,columns=96), 2021 (n=104,617, columns=96). The column of interest is the \_RFHLTH column (which represents categorical column for perceived health). These datasets are imbalanced and as such the recall metric shall be a point of focus in model evaluation. The profile reports will present detailed descriptive statistics about each column and its associated data. The tentative step by step Methodology from cleaning to the modeling is depicted **Figure 1 and 2**.

**Research Question 1:**

Dependent Variables: Perceived Health variable indicating subjective assessment of their overall health status, coded as ‘fair or poor health’ vs. ‘Good or Better Health’. All variables related to presence of cardiovascular diseases, such as Heart Attack or MI (Myocardial Infarction), Angina or CHD (coronary heart disease), and Stroke. These variables are coded as yes vs. no.

Independent Variables: Daily Dark Green Vegetable Consumption and Daily Fruit Consumption variables quantifies the daily intake of fruits and vegetables. Both coded as numerical values

**Research Question 2:**

Dependent Variables: Forgoing Medical Care due to Cost and Healthplan as the primary variables of interest were coded as yes vs. no.

Independent Variables: Race and ethnicity was included as white, black, native American, Asian, native Hawaiian/pacific islander, other race only, multiracial, and Hispanic. Sex variable categorizes individuals based on their gender and was coded as male and female. Income variable categorizes individuals based on their income levels, coded as Less than $15,000, $15,000 to $25,000, $25,000 to $35,000, $35,000 to $50,000, and more than $50,000. Education was coded as None, High School diploma, Attended College or Technical School, or Graduated from College or Technical School. Age was coded as 18-64 vs. 65 and older. Marital status was coded to include categories such as Married, Single, Widowed, Living Together, Divorced, and Separated. Employment status was also coded into categories, including Employed for Wages, Student, Self Employed, Homemaker, Retired, Unable, Out of Work for Less than 1 Year, and Out of Work for 1 Year or More.

**Research Question 3:**

Dependent Variables: All variables related to presence of chronic illnesses, such as Heart Attack or MI (Myocardial Infarction), Angina or CHD (coronary heart disease), Stroke, Diabetes, and depression. These variables are coded as yes vs. no.

Independent Variable: Exercise indicating whether the respondent exercised in the past month besides their regular job, was coded as yes vs. no.

**Research Question 4:**

Dependent Variables: All variables related to presence of chronic illnesses, such as Heart Attack or MI (Myocardial Infarction), Angina or CHD (coronary heart disease), Stroke, Diabetes, and depression. These variables are coded as yes vs. no.

Independent Variable: Perceived Health variable indicating subjective assessment of their overall health status, coded as ‘fair or poor health’ vs. ‘Good or Better Health’.

**Research Question 5:**

Dependent Variables: Binary class Perceived Health variable (RFHLTH) indicating subjective assessment of their overall health status, coded as ‘fair or poor health’ vs. ‘Good or Better Health’ and multiclass Perceived Health variable (GENHLTH) coded as poor, fair, good, very good and excellent.

Independent Variables: all variables and a selected subset of variables with the exception of the two above dependent variables.



**Figure 1: Diagram of Overall Methodology part a**

## Statistical Analyses:

Unpaired one-sided sample T-test was used to compare the mean fruit and vegetable consumptions between the two categories of the dichotomous variable, Perceived health. The unpaired one-sided sample T-test was also used to compare mean total CVD cases for individuals who consume more than 5 F&V versus those individuals that do not. Using pandas, a pivot table with each socioeconomic variable (race, sex, income, education, marital status, age and employment) and the corresponding percentage yes and no forgone medical care Response was generated. Chi-Squared Test for association was used to investigate the relationship between Chronic diseases and Exercise, as well as the relationship between Chronic diseases and perceived health. Logistic regression, Decision tree and Random Forrest model pipelines were used to model the dichotomous variable, Perceived health. Due to the imbalanced nature of our training dataset SMOTE (Synthetic Minority Over-sampling Technique) was used within each model pipeline as well as robust scaler to avoid data leakage by ensuring estimation only in the presence of the training data (Brownlee, 2020). Baseline model pipelines (all features) cross validation accuracy performance were compared with Model pipelines with selected features using Wilcoxon rank signed tests. Python 3.11.5 was used in all analyses. Python libraries were used to incorporate analytical tests. These include Pandas, Numpy, Seaborn, Matplotlib, Scikit-learn, Imbalanced-learn (Imblearn), Scipy. Our model's primary focus is on high recall to detect 'fair or poor health.' This prioritization enhances patient safety by ensuring timely medical interventions and reducing the chances of overlooking individuals in need. Simultaneously, our model's precision in identifying 'Good or Better Health' allows for a holistic approach to health management. This approach benefits individuals across health statuses, promoting preventive care and guaranteeing comprehensive care for all.



**Figure 2: Diagram of Overall Methodology part b**

## Findings and Interpretation:

**Research Question 1:**

A) Daily Dark Green Vegetable Consumption vs. Perceived Health:

Using an Unpaired One-sided Sample T-test with a confidence interval of 95%, we found a highly significant difference between mean daily dark green vegetable consumption for individuals with perceived good health and those with perceived bad health (p < 0.01). This significant difference indicates that those with good health consume significantly more dark green vegetables than those with bad health (p = 3.163e-112).

B) Daily Fruit Consumption vs. Perceived Health:

Similarly, an unpaired One-sided Sample T-test with a confidence interval of 95% revealed a highly significant difference between mean daily fruit consumption for individuals with perceived good health and those with perceived bad health (p < 0.01). This outcome suggests that individuals with good health have significantly higher daily fruit consumption than those with bad health (p = 3.323e-96).

C) Association between Total Fruit and Vegetable Consumption and Mean Total CVD Cases**:**

**Figure 3** presents an overview of the median Fruit and vegetable consumption for individuals with 1 to 4 total CVD (cardiovascular disease) cases. An Unpaired One-sided Sample T-test with a confidence interval of 95% was conducted to compare mean total CVD cases between individuals who consume fewer than 5 servings of fruits and vegetables and those who consume 5 or more servings. The results showed a highly significant difference (p < 0.01). This significant difference highlights that individuals with lower fruit and vegetable consumption have a higher mean total CVD cases (p = 1.287e-06). Overall, these analyses underscore the significant impact of dietary habits on perceived health and the prevalence of cardiovascular diseases, with p-values less than 0.01 at a 95% confidence level.

A graph of different colored boxes

Description automatically generated

**Figure 3: Total Fruit and vegetable consumption for individuals with 1 to 4 total CVD cases, 2019.**

**Research Question 2: Socioeconomic association with forgone medical care**

The analysis revealed that rates of forgoing medical care due to costs were highest among individuals who reported being out of work for less than a year and those with more extended periods of unemployment. Individuals with no formal education exhibited a notably high rate, approximately 22%. Similarly, those with income levels below $15,000 and between $15,000 to $25,000 displayed percentages as high as 22%. About 14% of adults aged 18 to 64 reported instances when they did not seek healthcare services due to cost constraints within the past 12 months. These findings are summarized in **Table 1**.

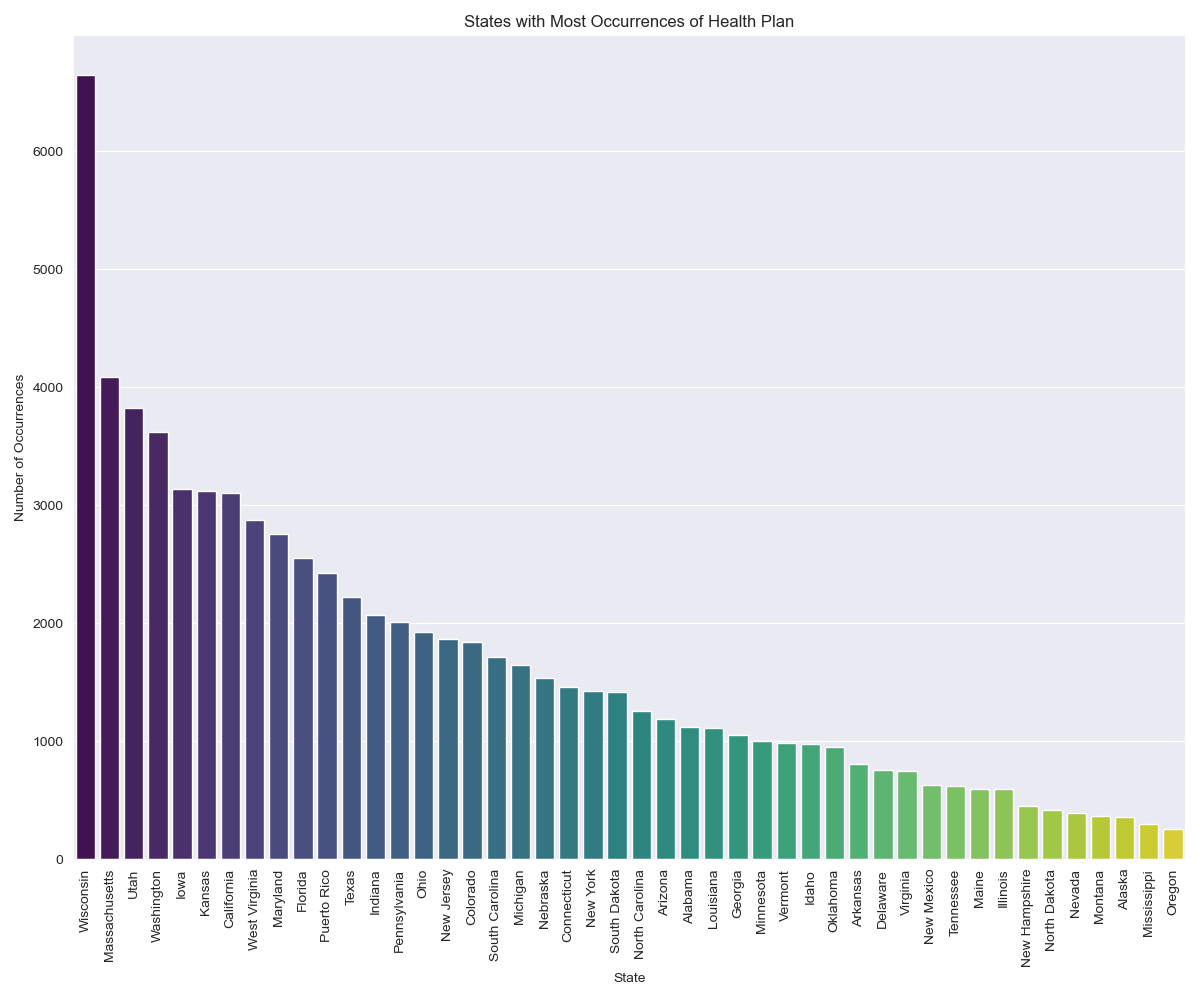
Furthermore, when healthcare access was examined in different regions of the United States, significant disparities were observed. Most Southern states tend to have a higher proportion of individuals without any Healthplan, while Northern states have higher occurrences of people with a form Healthplan. Wisconsin, in particular, stands out with the highest occurrences of both individuals with and without any Healthplan. **Figure 4** and **Figure 5** provide insights into the states with the highest and lowest occurrences of individuals with and without any Healthplan, respectively. Additionally**,** [**Figure 6**](file:///Users/obinnadinneya/Desktop/MY_BIGDATA_PROJECT/EDA/us_healthplan_map.html) illustrates the percentage of individuals with and without any Healthplan in each state.

These regional variations in healthcare access emphasize the importance of addressing healthcare disparities across the country. They underscore the need for targeted interventions and policy measures to ensure equitable access to healthcare services for all Americans.

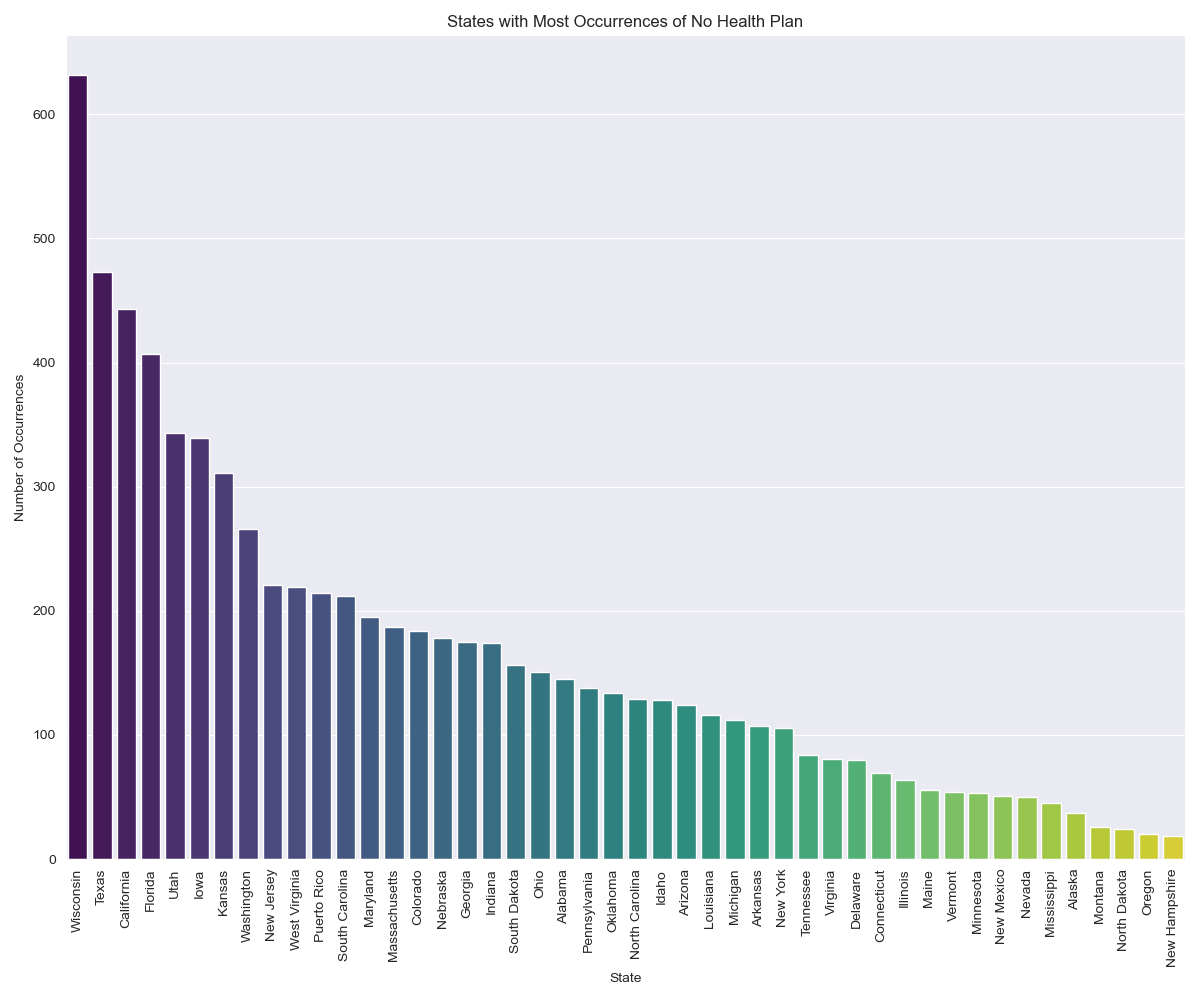
This analysis reveals the intricate relationship between socioeconomic factors and healthcare access, highlighting the challenges individuals encounter when seeking medical care. It underscores the need to address healthcare disparities for equitable outcomes.

****

**Table 1: Percentage Forgone medical Care by Race, Sex, Income level, Education level, Age, Marital status, and Employment Status**



**Figure 4: Percentage of individuals in each US State with any Health Plan, 2019**

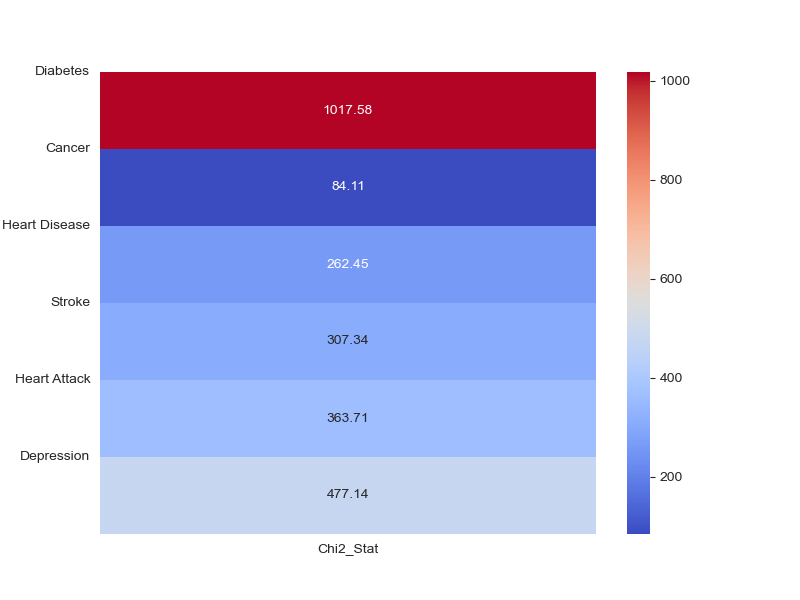


**Figure 5: Percentage of individuals in each US State without any Health Plan, 2019**

**Research Question 3: Association Between Exercise and Chronic Diseases**

**Figure 7** shows the chi-squared statistic for Exercise and each of the Chronic illnesses, with its strongest association being with diabetes. The direction of the association is explored by the contingency tables generated in the [EDA\_report](https://github.com/OBINNADINNEYA/BRFSS_BIGDATA_PROJECT/blob/branch_1/EDA/Exploratory%20Data%20Analysis.ipynb). Regular exercise appeared to be associated with a lower likelihood of diabetes, as individuals who engaged in physical activities had a notably lower incidence of diabetes (p < 0.01, 95% CI). A similar trend was observed for cancer, with non-exercisers exhibiting a higher prevalence (p < 0.01, 95% CI). Furthermore, the analysis uncovered associations between exercise and coronary heart disease and stroke .Those who refrained from regular exercise had a significantly higher risk of these conditions (p < 0.01, 95% CI).The impact of exercise extended to heart attacks, where non-exercisers had a significantly elevated risk (p < 0.01, 95% CI). On the mental health front, the absence of regular exercise was associated with a higher likelihood of depression, highlighting the importance of physical activity in promoting mental well-being (p < 0.01, 95% CI). These results are highly significant and suggest a reject of the null hypothesis that there is no association.

In summary, this analysis underscores the substantial role of regular physical exercise in reducing the risk of chronic diseases such as diabetes, cancer, coronary heart disease, stroke, and heart attacks. Additionally, it highlights the positive influence of exercise on mental health by lowering the likelihood of depression. It is essential to note that these findings represent measures of association and do not imply causation. Nonetheless, they emphasize the significance of advocating for physical activity as a preventive measure to enhance overall health and well-being.

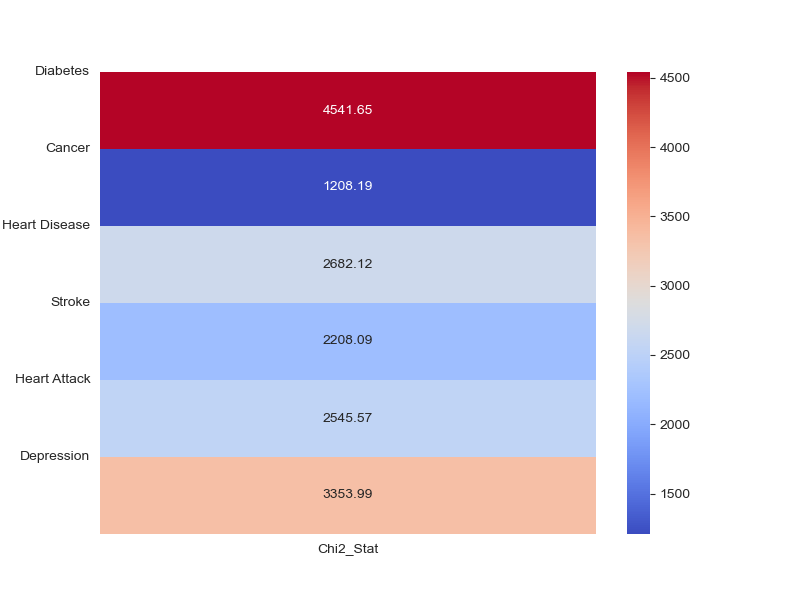
****

**Figure 7: Association Between Exercise and Chronic Diseases/ Heatmap of Chi-Squared Statistics**

**Research Question 4: Association Between Perceived Health and Chronic Diseases**

**Figure 8** shows the chi-squared statistic for Perceived health and each of the Chronic illnesses, with its strongest association being with diabetes. The direction of the association is explored by the contingency tables generated in the [EDA\_report](https://github.com/OBINNADINNEYA/BRFSS_BIGDATA_PROJECT/blob/branch_1/EDA/Exploratory%20Data%20Analysis.ipynb).Individuals who reported better perceived health had a significantly lower incidence of diabetes (p < 0.01, 95% CI). Similarly, better perceived health was associated with a lower prevalence of cancer (p < 0.01, 95% CI), coronary heart disease (p < 0.01, 95% CI), stroke (p < 0.01, 95% CI), and heart attacks (p < 0.01, 95% CI). Furthermore, we observed an association between perceived health status and depression. Those who reported better perceived health exhibited a lower likelihood of being diagnosed with depression (p < 0.01, 95% CI). These results are highly significant and suggest a reject of the null hypothesis that there is no association.

These findings underscore the importance of self-assessed health status as a potential indicator of overall well-being and risk of developing chronic diseases. It's important to note that these results indicate measures of association and do not imply causation. Nonetheless, they emphasize the significance of perceived health in understanding and potentially mitigating the risk of chronic diseases.

****

**Figure 8: Association Between Perceived Health and Chronic Diseases/ Heatmap of Chi-Squared Statistics.**

**Research Question 5: Accurate Prediction of Perceived Health using dietary habits, socioeconomic indicators, lifestyle choices, and individual metrics**

The impact of feature selection on model pipeline performance was assessed. The baseline model pipelines utilized all available features, encompassing a wide range of factors. In contrast, the Selected Feature Model pipelines employed a more streamlined approach, using only the 30 most important features obtained from an embedded selector (random forest feature selection). This reduction in feature complexity was undertaken to assess whether model performance could be improved while simplifying the model architecture. To assess whether feature selection had a statistically significant impact on model pipeline performance, we conducted Wilcoxon signed-rank tests for each model pipeline comparison with a confidence interval of 95%. The results of these tests showed p-values around 0.0625 for all model comparisons, which is above the threshold of 0.05. This indicates no statistically significant difference in accuracy scores between baseline and feature-selected model pipelines. Therefore, feature selection did not lead to a significant improvement or change in model performance. These results are tending towards statistical significance.

The Base Logistic Regression model was chosen based on its exceptional performance metrics. It achieved a high recall rate (79%) for 'fair or poor health' ,an overall accuracy of 84% and Matthews Correlation Coefficient (MCC) of 0.53. These metrics demonstrate the model's effectiveness in accurately identifying individuals with fair or poor health while maintaining a commendable level of overall accuracy. The efficiency of the selected Base Logistic Regression model was also assessed in terms of training and testing time. The training process took approximately 172 seconds, while the model executed predictions on the test data in just 0.05 seconds. This efficiency demonstrates that the model can be applied to new data in a timely manner, making it a practical choice for real-world applications. To evaluate the stability of the selected model, we conducted model validation using data from 2021. The results of this validation demonstrated remarkable consistency with the initial test results, indicating high stability with an accuracy of 85%, recall for ‘fair and poor health of 76% and Matthews Correlation Coefficient (MCC) of 0.52 (maintaining its predictive power while avoiding misclassification). Additionally, a multiclass classification was carried using GENHLTH as the variable of interest. This variable which consisted of 5 classes ( excellent, good, fair , poor, bad)

Was used to build predictive model pipelines using the same methodology as that for the binary classification. This resulted in overall poorly performing models, with low recall, overall Accuracy values and very long training times.

## Limitations:

The Behavioral Risk Factor Surveillance System (BRFSS) is a United States health-related telephone survey that collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. While the BRFSS is a rich data source for making health-related predictions, it has several limitations:

Self-Reported Data: BRFSS relies on individuals' self-reporting, which can lead to recall bias, under-reporting, or over-reporting of behaviors or conditions due to social desirability or memory issues.

Cross-Sectional Design: BRFSS is a cross-sectional survey, which captures a snapshot in time and cannot establish causality or account for changes over time in individuals' behaviors or health status.

Telephone Survey Limitations: The survey is conducted by telephone (landline and cell phone), which may exclude populations without access to phones or those who are less likely to participate in telephone surveys, like younger individuals who primarily use messaging apps.

Sampling Errors: Despite efforts to have a representative sample, there can be sampling errors, especially if response rates are low or if certain groups are underrepresented in the sample.

Non-Response Bias: The survey is subject to non-response bias; those who choose to respond may differ systematically from those who do not.

Uncontrolled Variations in Other Factors: When investigating the impact of socioeconomic factors on perceived health, it's important to note that variations in other influencing factors, such as dietary habits, lifestyle choices, and individual metrics, were not always controlled for in the analysis. These uncontrolled variations can introduce confounding factors that may affect the accuracy of predictions and the interpretation of results.

## Conclusion:

By prioritizing high recall for detecting 'fair or poor health,' our predictive model plays a pivotal role in patient safety and the provision of timely medical interventions. Minimizing false negatives is not just about accuracy; it's about the real-world impact on individuals who might otherwise be overlooked by the healthcare system. In parallel, our model's precision in identifying 'Good or Better Health' allows for a more nuanced, holistic approach to health management. This aspect of the model is crucial for designing proactive health interventions, promoting wellness, and optimizing resource allocation across the healthcare spectrum. The model's dual capability ensures that while we are vigilant in our efforts to support those in immediate need, we also empower those in 'Good or Better Health' to maintain and enhance their wellbeing. Such a holistic strategy enriches our healthcare system, fostering a preventive culture that benefits all strata of the population.

In summary, this approach establishes a healthcare environment that is preventive, responsive, and equitable, demonstrating a commitment to comprehensive care and well-being for every individual, irrespective of their current health status.

## References

Barreto, S. M., & Figueiredo, R. C. (2009). Chronic diseases, self-perceived health status and health risk behaviors: gender differences. Revista de Saúde Pública, 43(suppl 2), 38-47.

Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Publisher.

Carek, P. J., Laibstain, S. E., & Carek, S. M. (2011). Exercise for the treatment of depression and anxiety. International Journal of Psychiatry in Medicine, 41(1), 15-28.

Centers for Disease Control and Prevention. (2021). 2021 SMART: BRFSS City and County Data and Documentation. Retrieved from https://www.cdc.gov/brfss/smart/smart\_2021.html

Mbogori, T., & Arthur, T. M. (2019). Perception of Body Weight Status Is Associated with the Health and Food Intake Behaviors of Adolescents in the United States. American Journal of Lifestyle Medicine, 15(3). https://doi.org/10.1177/1559827619834507

Organisation for Economic Co-operation and Development (OECD). Health at a Glance 2021.

Palladino, R., Tayu Lee, J., Ashworth, M., Triassi, M., & Millett, C. (2016). Associations between multimorbidity, healthcare utilisation and health status: evidence from 16 European countries. Age and Ageing, 45(3), 431-435.

Pedersen, B. K., & Saltin, B. (2015). Exercise as medicine – evidence for prescribing exercise as therapy in 26 different chronic diseases. Scandinavian Journal of Medicine & Science in Sports, 25(S3), 1-72.

Towne Jr., S. D. (2017). Socioeconomic, Geospatial, and Geopolitical Disparities in Access to Health Care in the US 2011-2015. International Journal of Environmental Research and Public Health, 14(6), 573. doi: 10.3390/ijerph14060573.

Wallace, T. C., Bailey, R. L., Blumberg, J. B., Burton-Freeman, B., Chen, C. O., Crowe-White, K. M., Drewnowski, A., Hooshmand, S., Johnson, E., Lewis, R., Murray, R., Shapses, S. A., & Wang, D. D. (2020). Fruits, vegetables, and health: A comprehensive narrative, umbrella review of the science and recommendations for enhanced public policy to improve intake. Crit Rev Food Sci Nutr, 60(13), 2174-2211. doi: 10.1080/10408398.2019.1632258.

## 

## Github Links:

[Github BigData Project Repository](https://github.com/OBINNADINNEYA/BRFSS_BIGDATA_PROJECT)