Socioeconomic and Demographic Factors Influencing Self-Reported Health Outcomes

By Bridget Kim

Abstract/Executive Summary:

*This study explores how **socioeconomic and demographic factors** influence self-reported **health outcomes**, **mental health**, and **access to preventive care** in the U.S. using data from the **Behavioral Risk Factor Surveillance System (BRFSS)**. The BRFSS, overseen by the Centers for Disease Control and Prevention (CDC), collects annual data on health-related behaviors, chronic conditions, and use of preventive services among U.S. adults. This research focuses on key questions, such as the relationship between **income levels** and health status, variations in **physical activity** across demographics, and the impact of **health insurance** on access to care.*

The BRFSS dataset spans several decades (1993-2023), providing a unique opportunity to analyze trends over time. However, the evolution of data collection methods—from landline surveys to cell phone surveys—introduces potential biases, particularly toward more affluent populations in earlier years. To address these challenges, the study will clean and preprocess the data, ensuring that variables are comparable across years.

*The analysis will use both **regression** and **classification** techniques to understand the relationships between socioeconomic factors and health outcomes. By the end of the study, the goal is to identify key trends and correlations that can inform public health strategies and reduce health disparities across different communities.*

This paper presents the process of data cleaning, the methods for analysis, and the anticipated challenges in working with a large, multi-year dataset. Ultimately, the findings will provide insights into how socioeconomic and demographic factors shape health behaviors and outcomes in the U.S., helping to guide future health interventions and policies.

Introduction:

The **Behavioral Risk Factor Surveillance System (BRFSS)** is one of the largest health surveys in the world, overseen by the Centers for Disease Control and Prevention (CDC). Since 1984, it has been providing a continuous stream of data on the health-related behaviors, chronic health conditions, and use of preventive services among U.S. adults. This dataset, collected annually across all 50 states, Washington D.C., and U.S. territories, has become a cornerstone for shaping health policies, allocating resources, and informing public health initiatives across the country.

The primary objective of this project is to explore how **socioeconomic and demographic factors** influence self-reported **health outcomes**, **mental health**, and **access to preventive care** in the U.S. by utilizing the BRFSS dataset. Specifically, the analysis will address key research questions related to mental health, income and health status, physical activity variations by demographics, chronic conditions, and health insurance coverage. Understanding these relationships is critical in tackling health disparities and improving public health strategies that aim to reduce inequalities in health outcomes.

The BRFSS dataset spans several decades, providing a unique opportunity to track and analyze trends in health behaviors and outcomes over time. However, working with such an extensive dataset comes with its challenges. The data collection methods have evolved, from relying solely on landline phone surveys in the early years to including cell phone surveys since 2011. This shift in data collection methods introduces potential biases, as early surveys may have skewed toward more affluent populations. Additionally, the dataset spans multiple years, meaning that variable definitions, data formats, and categories may not always align perfectly. Thus, careful attention must be given to cleaning and harmonizing the data across years.

The analysis will begin by addressing smaller questions that contribute to the overall research question:

- **Self-reported rates of depression and anxiety** across different age groups (MENTHLTH).

- **Income levels** and their relationship with self-reported health status (INCOME2, GENHLTH).

- **Physical activity** differences based on **demographic factors** (EXERANY2, PASTAE1).

- **Chronic health conditions** and their correlation with socioeconomic factors (DIABETE3, CVDINFR4, ASTHMA3).

- **Health insurance status** and its impact on **access to preventive care** (HLTHPLN1).

These smaller questions will guide the analysis and provide insights into the larger issue of health disparities in the U.S. The ultimate goal is to provide evidence that can inform future public health interventions, focusing on how socioeconomic and demographic factors contribute to variations in health outcomes. Through this research, we hope to uncover the complex interactions between income, education, physical activity, mental health, and healthcare access—factors that shape the health landscape for different communities across the U.S.

In this submission, we will focus on cleaning and preparing the data for analysis, addressing challenges such as missing data, variable standardization, and outlier detection. Once the data is cleaned, we will proceed to analyze the relationships between the identified variables, providing both descriptive and statistical insights to answer the research questions. By the end of the study, we aim to contribute meaningful findings that can guide future health interventions and policies.

Data Overview:

Behavioral Risk Factor Surveillance System (BRFSS)

The **Behavioral Risk Factor Surveillance System (BRFSS)** is the largest continuously conducted health survey system in the world, overseen by the **Centers for Disease Control and Prevention (CDC)**. It collects data from U.S. adults (ages 18+) about their health-related risk behaviors, chronic health conditions, and use of preventive services. This dataset spans multiple years, including data from 1993 to 2023, providing insights into trends in health behaviors and outcomes.

The BRFSS is unique in that it collects data annually in all 50 states, Washington D.C., and U.S. territories. It is used by public health professionals and policymakers to allocate resources, track health disparities, and design effective interventions. The system is a critical tool for tracking national and state-level trends in health behaviors, which can significantly inform public health programming.

Core Purpose

Since its inception, the BRFSS has aimed to:

- **Monitor health behaviors** contributing to the leading causes of death and disability.
- **Track access to healthcare** and use of preventive services.
- **Provide timely, localized health data** to support public health planning and evaluation.

Significance and Use

The data collected by BRFSS are used by various stakeholders, including:

- **Public health agencies**, who use the data to allocate resources and design targeted interventions.

- **Researchers and policymakers**, who analyze trends to understand and address disparities in health outcomes.
- **Local governments**, who tailor community health initiatives based on real-time behavioral data.

Key Topics Included in the Dataset:

The BRFSS dataset includes information on various aspects of health and lifestyle, such as:

- **Healthcare access** (health insurance status, use of preventive care)
- **Tobacco and alcohol use**
- **Physical activity** and exercise (EXERANY2, PASTAE1)
- **Diet and nutrition** (e.g., fruit and vegetable consumption, FRUITE1, VEGESU1)
- **Hypertension and cholesterol** management
- **HIV/AIDS knowledge and prevention**
- **Cancer screening** practices
- **Immunizations**
- **Injury prevention**

How BRFSS Has Evolved Over Time

- **Topics Covered**:
    - In earlier years (e.g., 1999), topics mainly included **smoking**, **alcohol use**, **exercise**, and **screenings**. As the dataset evolved, it began to incorporate more comprehensive topics, such as **mental health**, **insurance access**, **social determinants of health**, **opioids**, and **COVID-19**.
- **Data Collection Method**:
    - Initially, the survey relied on **landline telephone surveys**. Since 2011, it expanded to include **cell phone surveys** to ensure broader representation.

- **Weighting Method**:
    - Early data collection used **post-stratification**, but over time, the method was improved to **raking (iterative proportional fitting)** for better accuracy in demographic representation.
- **Customization**:
    - The BRFSS began with a **standard core set of questions**, with optional modules included over time. This customization allows for state-specific data collection and has improved the flexibility of the survey to capture emerging public health issues.
- **Equity Focus**:
    - While the early years had limited demographic analysis, recent years emphasize **health equity**, focusing on factors like **race**, **income**, **education**, and **geography**.

Plan for Cleaning:

The data from the BRFSS is **cleaned** to ensure it is usable for analysis, especially considering the varying data collection methods and possible biases in earlier years.

Data Challenges

1. **Skew Towards More Affluent Populations:**
    - Data from earlier years (e.g., 1993) may be skewed towards more affluent populations, particularly because landline surveys in those years did not reach low-income households as effectively as cell phone surveys can today.
2. **Consistency in Variable Definitions:**
    - One of the key challenges is ensuring that variables across different years (e.g., 1993 to 2023) are defined consistently. This includes checking the value definitions, particularly when variables have evolved over time (e.g., the introduction of cell phone surveys since 2011).
3. **Missing Data Handling:**

   - Missing values are a common issue, especially across different years. Imputation methods or exclusion of rows/columns with significant amounts of missing data may be necessary. Careful attention must be given to avoid introducing bias when handling missing data.

4. **Outliers:**
   - Outliers, especially in income or other continuous variables (e.g., extremely high reported income), may need to be capped or removed.

5. **Data Transformation:**
   - Some variables may need to be transformed or re-coded to match their definitions across years. For instance, categorical variables might need to be standardized to facilitate comparison.

Next Steps:

Consolidate Variables:

- A major step is to **compile** the relevant variables across all years into a single file. This means aligning each year by variable names and value definitions, ensuring consistency in how data is recorded (e.g., recoding health-related variables to standard definitions).

Clean and Preprocess:

- **Data Imputation**: Missing values will be handled by imputation (e.g., median for continuous variables, mode for categorical variables).

- **Outlier Detection**: Identify and handle outliers, especially for continuous variables like **income** and **age**.

- **Variable Transformation**: Adjust or create new features that might be necessary, such as categorizing income ranges or creating interaction terms between age and income for predictive modeling.

Descriptive Statistics:

- **Summary Statistics**: Generate basic summary statistics (mean, median, standard deviation, etc.) for all variables to better understand data distributions and identify any potential anomalies.

- **Correlation Analysis**: Examine correlations between **socioeconomic variables** (e.g., income, education) and **health outcomes** to identify relationships worth exploring in predictive models.

Visualizations:

- Create visualizations to compare health outcomes, mental health, and access to care across different **demographic groups** (e.g., age, income, gender). For example, a box plot for income distribution across age groups or a heatmap to visualize correlations between health behaviors and outcomes.

Methods

1. Observation in the Study:

Each observation in the dataset represents a single survey respondent in the **Behavioral Risk Factor Surveillance System (BRFSS)**. The data is self-reported and includes a variety of health behaviors, conditions, and demographic characteristics such as age, gender, income, physical activity, mental health status, and access to healthcare services. These responses, gathered from 1993 to 2023, form the basis of the analysis.

2. Type of Learning (Supervised vs. Unsupervised):

The study will primarily use **supervised learning** techniques since the goal is to predict outcomes based on a set of input features. Specifically:

- **Regression:** We will use regression techniques to predict continuous variables, such as general health status or mental health levels (e.g., depression or anxiety).

- **Classification:** We will also use classification methods to predict binary outcomes, such as the presence of chronic conditions or mental health disorders.

3. Models and Algorithms:

The following models and algorithms will be used to answer the research questions:

- **Linear Regression**: Initially, linear regression will be used to model the relationships between socioeconomic factors (such as income, education, and occupation) and self-reported health outcomes (e.g., general health status, mental health).

- **Logistic Regression and Decision Trees**: These models will be applied for classification tasks, such as predicting whether an individual is likely to report certain chronic conditions (e.g., asthma, diabetes) or mental health disorders. Depending on performance, we will choose the best model for these binary outcomes.

- **Principal Component Analysis (PCA)**: If the dataset contains highly correlated features, PCA will be applied for dimensionality reduction to simplify the model without losing critical information.

- **Random Forests and LASSO**: To handle complex interactions and multicollinearity among features, random forests and LASSO (Least Absolute Shrinkage and Selection Operator) will be explored to improve predictive accuracy and enhance the robustness of the model.

4. Evaluation of the Model's Success:

The success of the models will be evaluated based on several performance metrics:

- **For Regression Models:**
  - **R^2 (Coefficient of Determination)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
  - **RMSE (Root Mean Squared Error)**: A measure of the average magnitude of the errors between predicted and observed values.
- **For Classification Models:**
  - **Accuracy**: The percentage of correct predictions.
  - **F1 Score**: A measure of a model's accuracy, balancing precision and recall.
  - **Sensitivity and Specificity**: Sensitivity measures how well the model detects positive cases, while specificity measures how well it detects negative cases.
- **Cross-Validation**: To assess how well the model generalizes to unseen data, **cross-validation** (with 5 folds) will be used. This helps in understanding the model's stability and potential overfitting to the training set.

5. Potential Issues and How to Address Them:

- **Missing Data**: There may be missing values for certain years or variables. Missing data will be addressed either through **imputation** (using median for numeric variables and mode for categorical ones) or **exclusion** of rows/columns with excessive missing data.

- **Data Bias**: Earlier BRFSS data, especially from years like 1993, may be biased due to limited phone access (only landline surveys). This may skew the data towards more affluent populations. We can mitigate this by **weighting** the data to account for demographic imbalances, or adjusting for socio-economic variables like income and education in the models.

- **Multicollinearity**: If predictor variables are highly correlated, this could destabilize the regression models. To mitigate this, we will use techniques such as **LASSO** for regularization or **PCA** for dimensionality reduction.

- **Model Overfitting**: Overfitting can occur if the model becomes too complex. This will be controlled through the use of **cross-validation**, **regularization**, and careful feature selection.

6. Feature Engineering:

- **Encoding Categorical Variables**: Categorical variables (e.g., gender, region) will be **one-hot encoded** to transform them into a format suitable for regression and classification models.

- **Scaling**: Continuous variables (e.g., age, income, physical activity) will be **scaled** using techniques such as **Min-Max scaling** or **Standard Scaling** to ensure all features are on the same scale, which is important for certain models like logistic regression and decision trees.

7. Results Communication:

- **Regression Models**: Results will be communicated using **tables of coefficients** to explain the relationship between predictors (socio-economic variables) and the target health outcomes.

- **Classification Models**: Results for classification models will be presented through **confusion matrices**, showing how many predictions were correct (true positives/negatives) and how many were incorrect (false positives/negatives).

By addressing these steps, we will analyze the relationships between socioeconomic factors and self-reported health outcomes and provide insights that may inform public health policies and interventions.

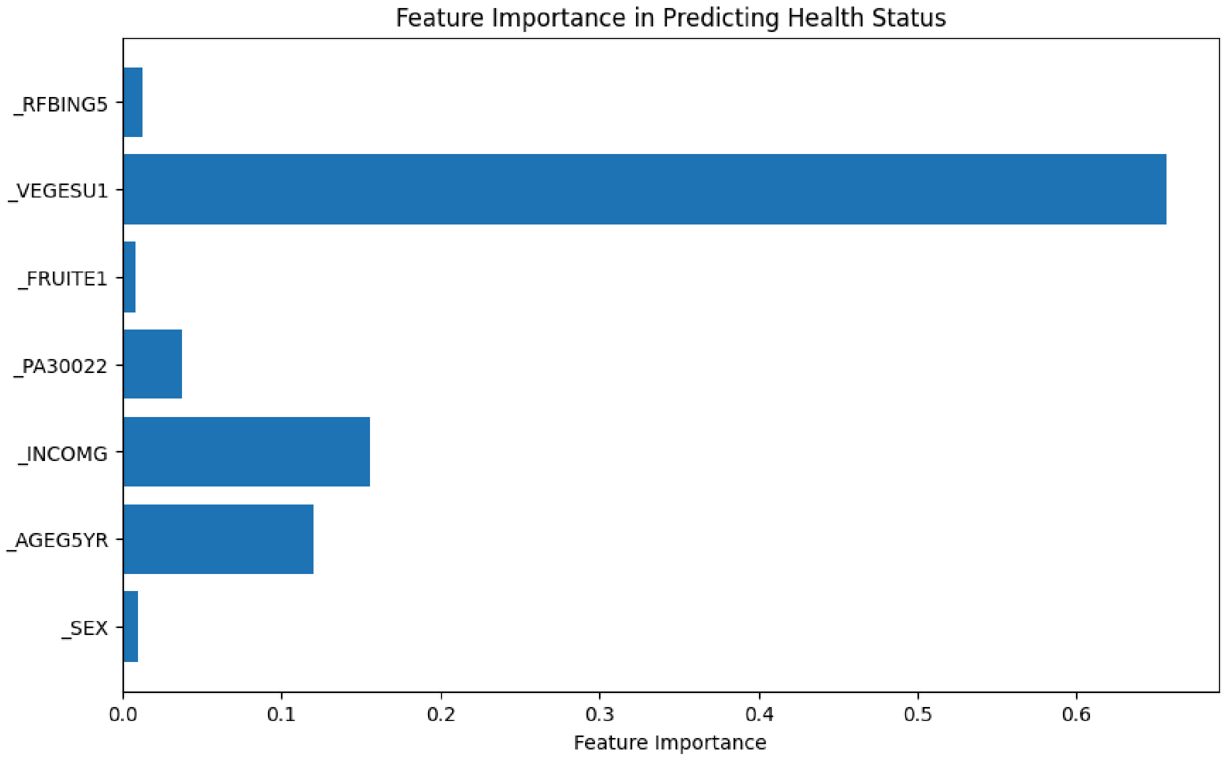
Results

Model Performance

The model’s overall performance in predicting health status for 2019 was evaluated using a Random Forest Classifier. The results yielded an **accuracy score of 99.75%**, which initially seems impressive. However, a closer look at the **classification report** revealed an issue: the model performed exceptionally well on the majority classes (1.0 for "good health" and 2.0 for "fair health") but struggled with the minority class (9.0, which represents unknown health status). Specifically, the recall for the class 9.0 was only 9%, which indicates the model's poor ability to predict the "unknown" health status accurately.

The **macro average** of the classification metrics (precision, recall, and F1-score) was much lower than the **weighted average**, which highlighted the class imbalance. The disparity between these averages suggested that the model’s performance was dominated by the majority classes, and it struggled to correctly classify the minority class.

Feature Importance



The feature importance plot from the Random Forest model revealed the following insights:

1. **Vegetable Consumption (`_VEGESU1`)** emerged as the most influential feature in predicting health status, accounting for over 60% of the model's predictive power. This suggests that **vegetable consumption** plays a central role in determining overall health status.

2. **Binge Drinking Behavior (`_RFBING5`)** was the second most important feature, which indicates that **excessive alcohol consumption** has a strong association with poorer health outcomes.

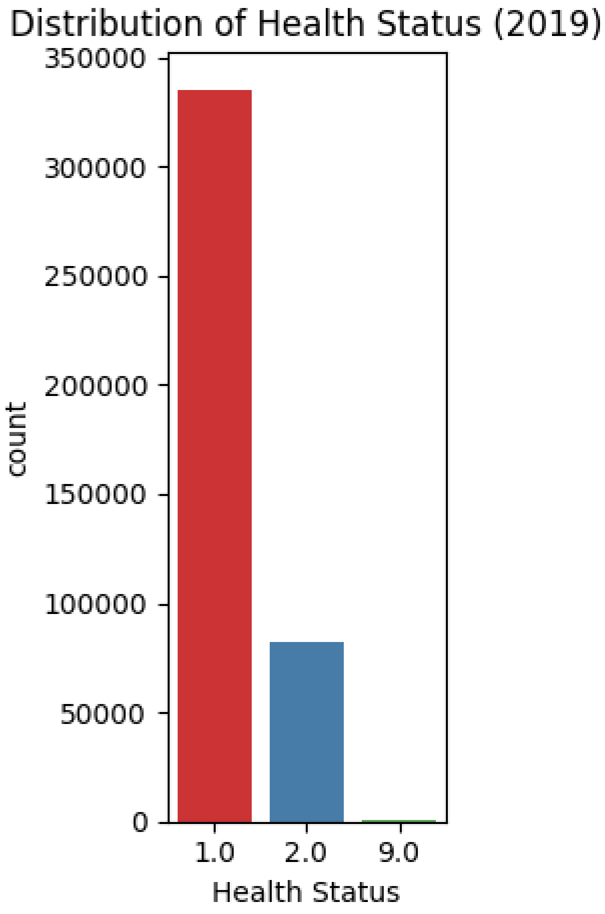
3. **Fruit Consumption (`_FRUITE1`)** also contributed to the model's predictions, but with lower importance compared to vegetable consumption. This suggests that both **fruit and vegetable consumption** are relevant for health, although vegetable intake had a stronger impact.

4. **Physical Activity (`_PA30022`)** showed moderate importance in predicting health status, with higher levels of physical activity generally being associated with better health outcomes.

5. **Income Group (`_INCOMG`)** and **Age Group (`_AGEG5YR`)** were also important but less so than lifestyle-related features. This indicates that **socioeconomic factors** like income do influence health, but lifestyle choices like diet and physical activity are more strongly predictive of health outcomes.

6. **Gender (`_SEX`)** was the least important feature, suggesting that **sex** had minimal impact on health status compared to lifestyle factors and socioeconomic status in this dataset.

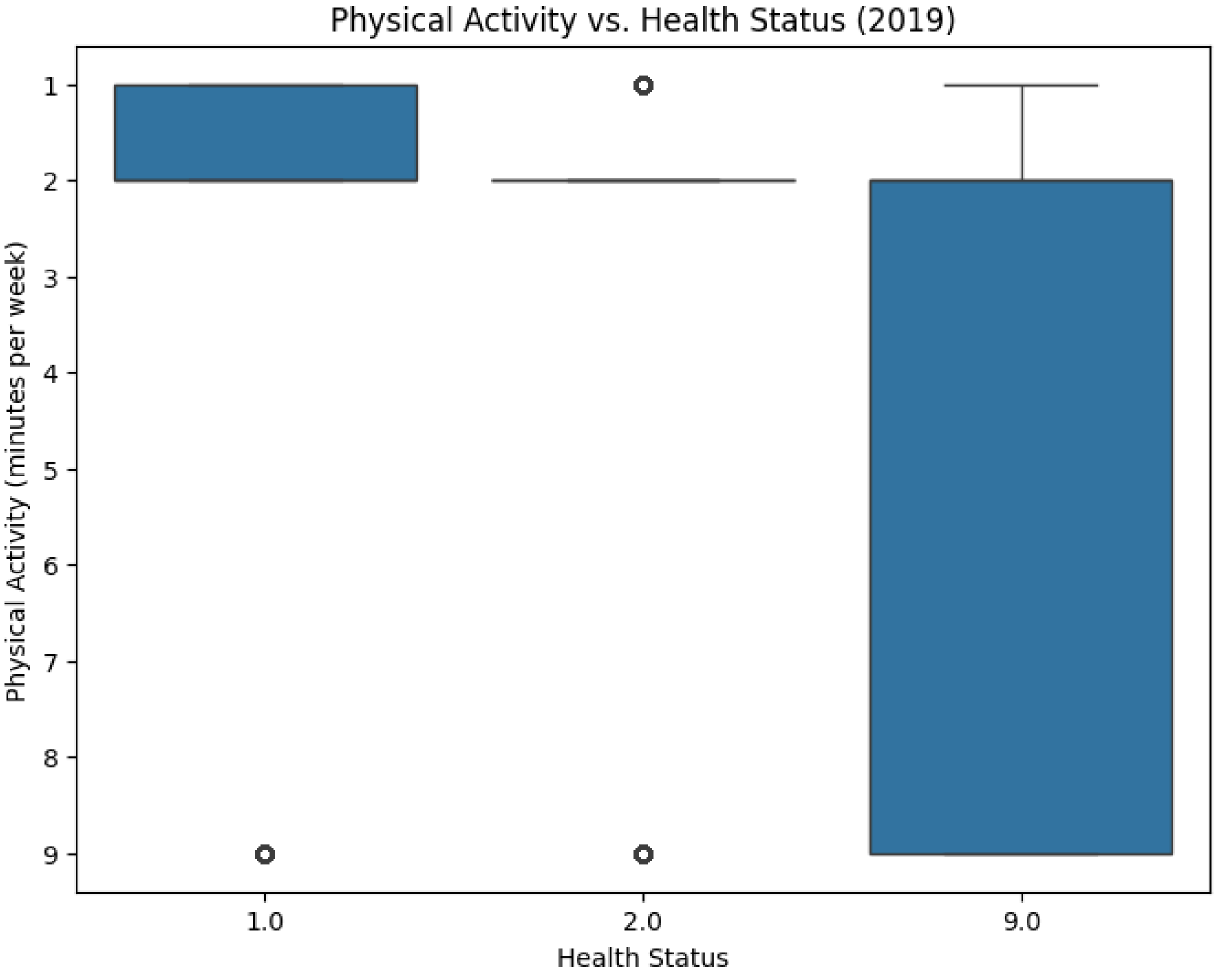
Health Status Distribution (2019)



The **distribution of health status** for 2019 showed that the vast majority of respondents reported being in **good health** (1.0), followed by a smaller proportion in **fair health** (2.0). Very few respondents were categorized as **having unknown health status** (9.0). This skew in the data contributed to a **class imbalance** in the model, with the model heavily predicting "good health" (1.0) while underperforming on the minority class (9.0).

The imbalance between the classes implies that the model is more likely to classify most individuals as being in **good health**, which is not ideal for accurately predicting all categories. This issue can be mitigated by adjusting for class imbalance through techniques such as **oversampling the minority class** or **undersampling the majority class**.

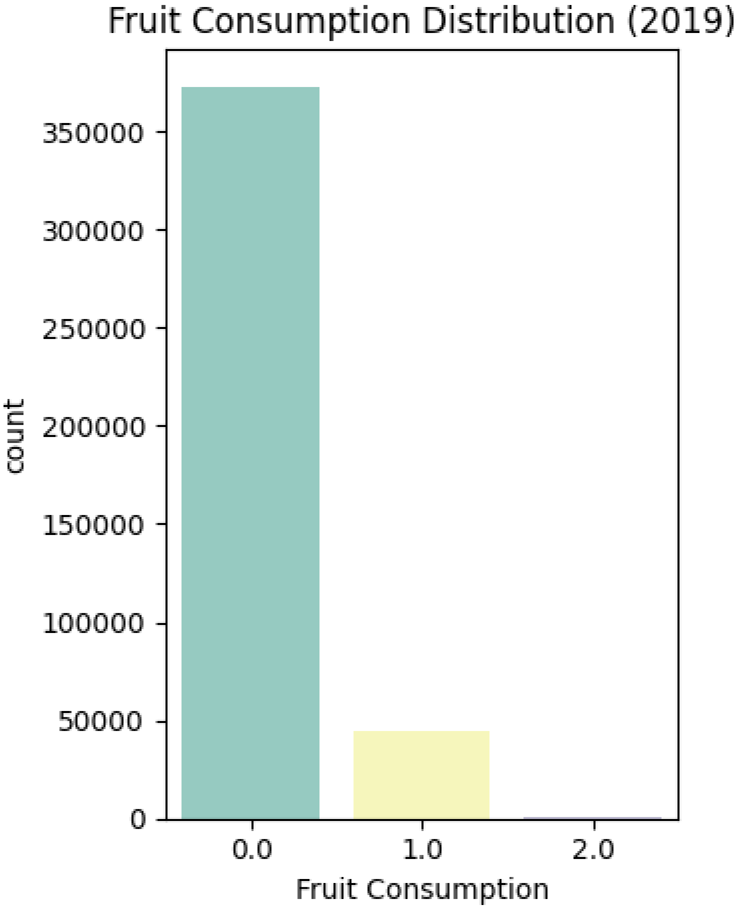
Physical Activity vs. Health Status (2019)



A boxplot visualizing **physical activity** (measured in minutes per week) against health status revealed an interesting trend: respondents reporting **better health** (1.0) tended to have low to moderate levels of physical activity, while those in **poorer health** (9.0) had higher physical activity levels. This inverse relationship suggests that, for certain individuals, **physical activity** might be a compensatory factor for **poor health status**, especially for those with underlying chronic conditions or mental health issues.

This finding is consistent with the notion that physical activity is a **strong predictor** of health. However, the **outlier** values (with higher physical activity and worse health) suggest that factors beyond just physical activity, such as **mental health** or chronic conditions, could significantly affect health status.

Fruit Consumption Distribution (2019)



The distribution of **fruit consumption** (measured on a scale) was highly skewed, with most individuals reporting **minimal or no fruit consumption** (0). This could indicate a **lack of dietary diversity** or possibly reflect a broader trend of individuals prioritizing other food groups, like vegetables or proteins, in their diets.

Despite its relatively lower importance in predicting health status (compared to vegetable consumption), fruit consumption still plays a role in overall health. The limited consumption observed in the dataset might highlight an area of focus for improving public health outcomes, suggesting that efforts to increase fruit consumption could positively influence general health.

Insights and Limitations

The results highlight that **lifestyle factors**, particularly **vegetable consumption** and **binge drinking behavior**, are the most significant predictors of health status. However, **socioeconomic factors** like **income** and **age** also contribute to health outcomes, albeit to a lesser extent. The model's ability to predict health status was affected by **class imbalance**, which could be addressed in future work by focusing on improving the model's performance for minority classes like "fair health" and "unknown health."

The results also underscore the need for further exploration of **physical activity** and **dietary habits** in predicting health. While physical activity showed strong associations with health outcomes, **fruit and vegetable consumption** remained influential in determining health status, reinforcing the importance of a balanced diet.

Finally, future work could address the **class imbalance**, explore interactions between **socioeconomic and lifestyle factors**, and extend the analysis to other years of data to capture trends over time.

Conclusion

The analysis of the Behavioral Risk Factor Surveillance System (BRFSS) data from 2019 reveals significant insights into the relationship between **socioeconomic, demographic, and lifestyle factors** and **self-reported health outcomes**. The Random Forest Classifier model showed high accuracy (99.75%) but also revealed a **class imbalance** issue, particularly in predicting **"unknown health status"** (9.0), which the model struggled to classify correctly. The majority class, "good health" (1.0), dominated the predictions, overshadowing the fair health (2.0) and unknown categories. This imbalance suggests that while the model can predict health status for the larger, dominant class, it requires further refinement to improve prediction accuracy for minority classes.

Key variables such as **vegetable consumption (_VEGESU1)**, **binge drinking behavior (_RFBING5)**, and **physical activity (_PA30022)** emerged as the most significant predictors of health status. These findings reinforce the idea that **lifestyle factors**, especially diet and physical activity, play a critical role in determining health outcomes. **Income** and **age**, while contributing to health outcomes, had less influence compared to lifestyle habits. The model also highlighted that **gender (_SEX)** was the least important variable, suggesting it did not significantly impact health outcomes once other factors were considered.

Criticisms and Limitations

While the results provide valuable insights, several challenges and limitations were encountered during the analysis:

1. **Class Imbalance**: The dominant "good health" category led to a skewed model performance. The model's poor performance in predicting the minority class (9.0, unknown health status) indicates that a **balanced approach** (such as oversampling or undersampling) is necessary to improve predictions across all health status categories.

2. **Data Inconsistency**: The data from different years exhibited **variability in variable definitions**, especially between earlier years (e.g., 1993) and more recent years. This inconsistency made it difficult to compare trends over time, and the lack of **documentation** for early years further complicated the task of aligning variables. Due to this, the analysis was limited to the 2019 dataset, forgoing the originally intended comparative analysis across years.

3. **Missing Data and Data Quality**: Some **missing data** in certain years (particularly in the 2023 dataset) hindered the full exploration of trends across multiple years. Additionally, attempts to convert the 2023 dataset were unsuccessful, which led to a **reduced scope** of analysis and potentially overlooked insights from the most recent data.

4. **Overcomplicating the Model**: Including too many features, especially ones with low importance, made the model harder to interpret. The model's high complexity, driven by **many redundant variables**, led to difficulty in understanding the individual impact of each feature. Moving forward, focusing on key variables with **strong predictive power**, like **vegetable consumption, binge drinking, and physical activity**, will likely lead to more interpretable and actionable results.

5. **Variable Issues**: Some variables, like **fruit consumption (_FRUITE1)**, showed lower importance, yet their impact was not fully explored due to the broad range of features considered. In future iterations, **targeted feature engineering** and **dimension reduction techniques** such as PCA could be used to focus on more meaningful predictors of health status.

Moving Forward

To improve model performance and interpretation, the next steps should focus on addressing the **class imbalance**, refining the **selection of key features**, and exploring **time-based trends** across different years. Additionally, more effort should be put into cleaning the data from **earlier years** (e.g., 1993) to align variables and ensure comparability over time.

By refining the model and expanding the analysis to other years, a clearer picture of the evolving relationship between **socioeconomic factors, lifestyle choices**, and **health outcomes** can emerge, aiding in public health interventions aimed at reducing disparities and improving overall well-being.

References/Bibliography:

1. Centers for Disease Control and Prevention. (n.d.). *BRFSS Annual Data.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_data.htm
2. Centers for Disease Control and Prevention. (1993). *BRFSS Annual Data 1993.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_1993.htm
3. Centers for Disease Control and Prevention. (1999). *BRFSS Annual Data 1999.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_1999.htm
4. Centers for Disease Control and Prevention. (2003). *BRFSS Annual Data 2003.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2003.htm
5. Centers for Disease Control and Prevention. (2009). *BRFSS Annual Data 2009.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2009.htm
6. Centers for Disease Control and Prevention. (2013). *BRFSS Annual Data 2013.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2013.html
7. Centers for Disease Control and Prevention. (2019). *BRFSS Annual Data 2019.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2019.html
8. Centers for Disease Control and Prevention. (2023). *BRFSS Annual Data 2023.* Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2023.html