Team 3: (Dhanush Mathivanan, Vaishali Jayaram, Simbarashe Mpofu)
DATS 6103: Summary Report
Professor Ning Rui
November , 2025

Financial Factors influence on general health

Introduction

In the United States, income inequality remains one of the strongest predictors of differences in health outcomes across communities. Using the 2024 Behavioral Risk Factor Surveillance System (BRFSS), one of the largest ongoing health surveys in the country, we examined how income related factors influence an individual's likelihood of reporting poor health. Our analysis shows that financial stressors including income level, employment status, insurance coverage, and the ability to afford medical care play an equally substantial role in shaping health outcomes.

For this project, we focused on identifying which social and economic determinants of health contribute most strongly to poor health and whether these factors can be used to accurately predict a person's reported health status. We also studied how additional lifestyle and demographic variables (such as age, sex, BMI, and chronic conditions) interact with income to influence health.

Finally, to evaluate how well these factors can predict health outcomes, we developed multiple machine-learning models including Logistic Regression, Random Forest, and CatBoost. While all models performed reasonably well, CatBoost emerged as the best-performing model, demonstrating strong accuracy and balanced recall when identifying individuals in poor health. This suggests that a combination of income-related variables and health-behavior factors provides meaningful insight into understanding and predicting health disparities across U.S. communities.

SMART Questions

With To guide our analysis and ensure our project addressed meaningful public-health issues, we developed research questions using the 2024 BRFSS survey data:

1. Which financial and socioeconomic factors most significantly predict whether U.S. adults report poor general health.

Literature Review

Previous research consistently shows that social and economic conditions such as income, education, employment, and access to healthcare play a major role in shaping people's health. Reports from the CDC (2023) and the World Health Organization (2024) highlight that these social determinants often influence health outcomes more than medical care itself. People with lower income or financial stress are more likely to experience chronic illnesses, mental health challenges, and difficulties accessing regular medical care.

Studies using national surveys like BRFSS also find that individuals with lower socioeconomic status have significantly higher odds of reporting fair or poor health across the United States. Broader public-health work has

shown that financial strain, food insecurity, and limited insurance coverage all contribute to worse physical and mental health (National Academies of Sciences, Engineering, and Medicine, 2019).

Our project builds on this research by using the newest 2024 BRFSS dataset to examine how income-related factors, together with lifestyle and disability indicators, influence health outcomes. We also compare different predictive models to understand which socioeconomic factors are the strongest predictors of poor health.

Description Of Data

Our dataset comes from the 2024 Behavioral Risk Factor Surveillance System (BRFSS), a nationwide health survey conducted by the CDC. The combined dataset includes responses from 49 states, the District of Columbia**,** Guam**,** Puerto Rico, and the U.S. Virgin Islands. In total, the dataset contains 457,670 records**,** each representing an adult respondent with detailed information on demographics, income, health behaviors, and chronic conditions.

Data Cleaning and Preparation of Data

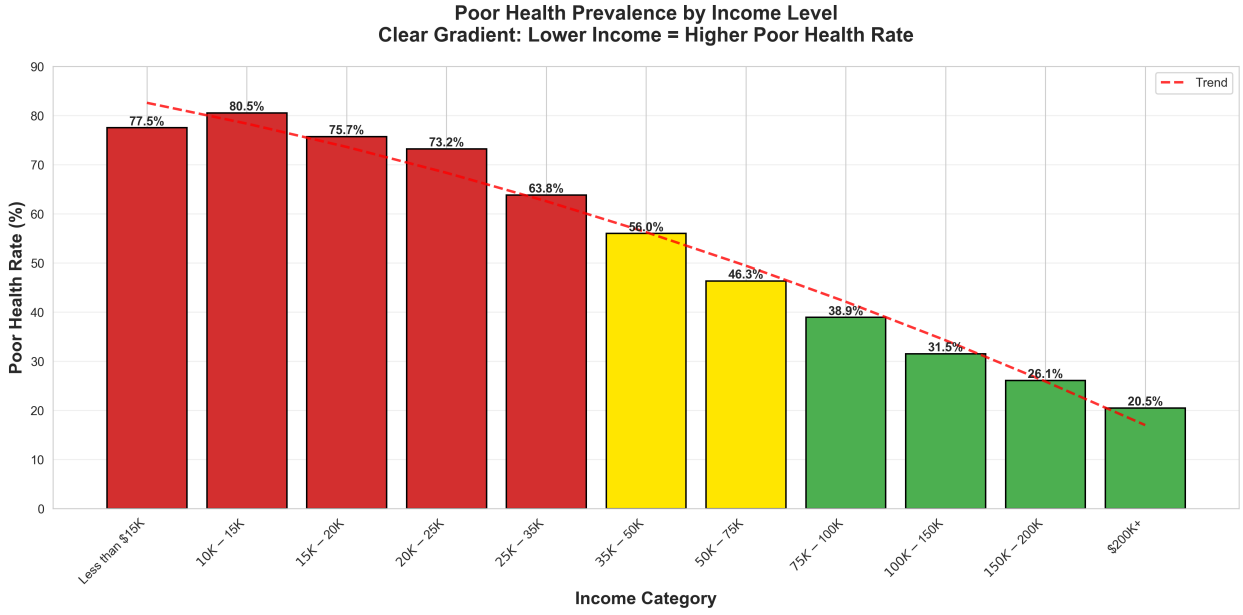
To prepare the BRFSS dataset for analysis, we applied extensive cleaning based on the survey codebook, which uses special codes (7, 9, 77, 88, 99) to denote "Don't know," "Refused," or missing responses. These invalid codes were removed across all key variables, including Income, Insurance, Employment, Education, Exercise, Chronic Conditions, and Disability indicators. The largest reduction came from the Income_Categories variable, where 78,153 invalid records (codes 77/99) were dropped. After applying variable-specific cleaning to all fields, the dataset was reduced from 457,670 records to 256,684 valid responses. We then removed implausible BMI values (<12 or >60), resulting in 242,948 clean records.

Next, we created our binary outcome variable (Poor_Health) and selected 17 socioeconomic, lifestyle, disability, and clinical features for modelling. Missing numeric values were imputed using medians. Because poor health was underrepresented in the sample (only 17%), we applied a state-based under sampling method that evenly matched each state's poor-health cases with an equal number of good-health cases. This produced a fully balanced dataset consisting of 41,916 good-health and 41,916 poor-health records, yielding a final modelling dataset of 83,832 cleaned and balanced observations.

EDA and Graphs

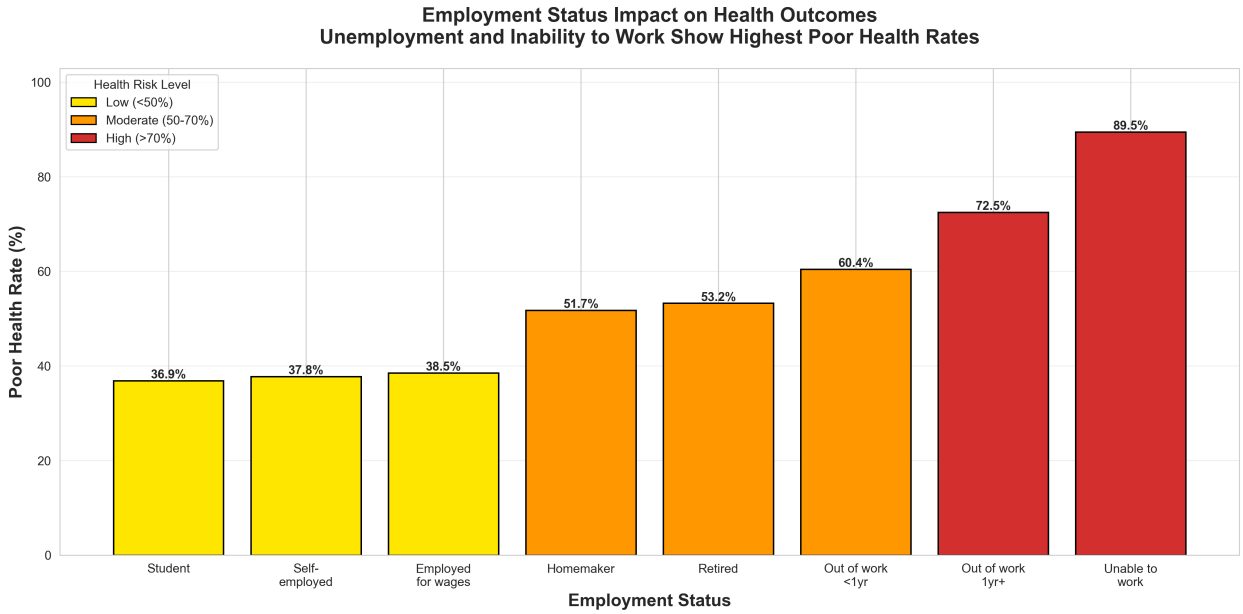
To better understand how income and financial stressors relate to health outcomes, we conducted exploratory data analysis using the cleaned BRFSS dataset. Several clear patterns emerged across income, employment status, and healthcare affordability. Overall, the EDA consistently shows that low income is strongly associated with higher rates of poor health**.**

Figure 1. Poor Health Prevalence by Income Category



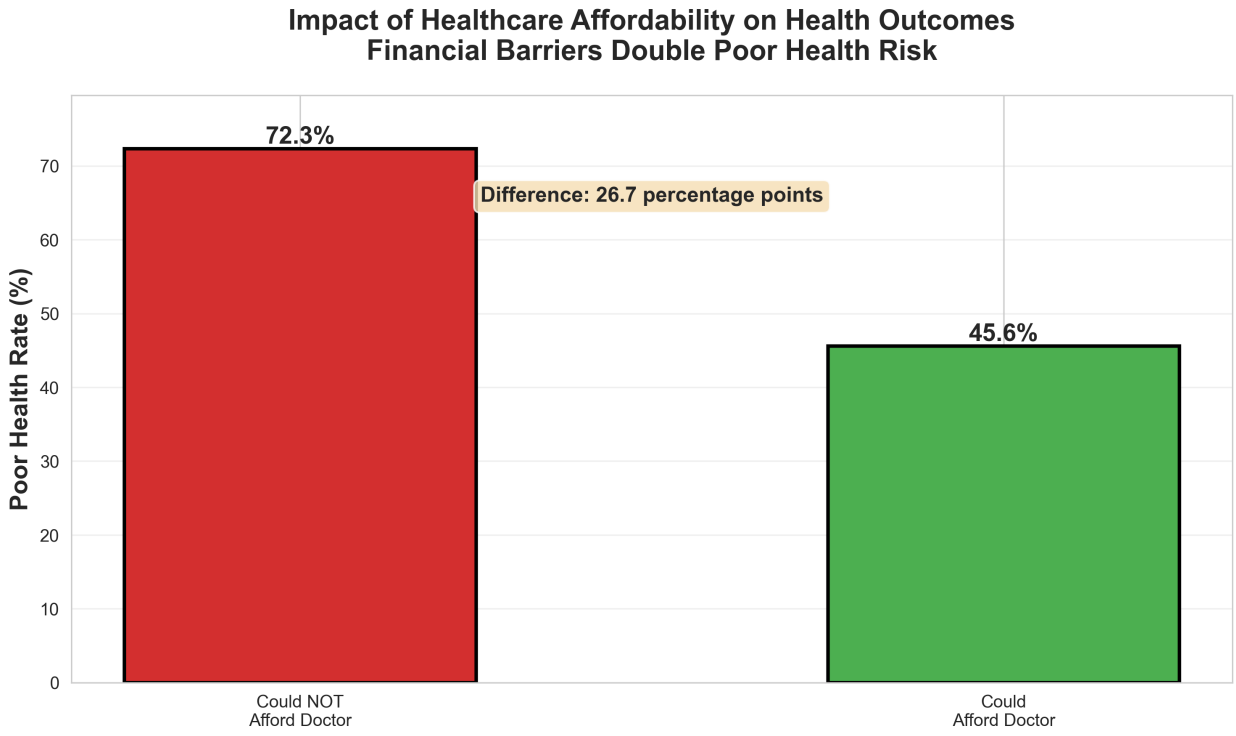
Lower income levels are associated with substantially higher poor-health rates, with the lowest income group exceeding 75% prevalence. The trend line illustrates the clear downward gradient as income increases.

Figure 2. Employment Status and Poor Health Rates



The highest poor-health rates occur among individuals “Unable to work” (89.5%) and those “Out of work for 1+ years” (72.5%). In contrast, employed individuals and students have the lowest rates.

Figure 3. Poor Health Rates by Ability to Afford Medical Care



Individuals who could not afford a doctor had a poor-health rate of 72.3%, compared with only 45.6% among those who could access care

Modelling

The goal of this analysis was to build a reliable predictive model that identifies whether an individual reports poor health(1*)* or good health(0*)*, based on financial, socioeconomic, behavioral, and health-related factors. Because the outcome is binary, we used three classification models that is **Logistic Regression**, **Random Forest**, and **CatBoost** model and evaluated their performance after balancing the data (50% good health, 50% poor health). The main metrics of interest were accuracy, recall, and AUC-ROC, with special emphasis on correctly identifying those in poor health.

Model Comparison

Model	Accuracy	AUC-ROC	Recall (Poor Health)	F1 (Poor Health)
Logistic Regression	0.7598	0.8418	0.73	0.75
Random Forest	0.7752	0.8549	0.78	0.78
CatBoost	**0.7817**	**0.8614**	**0.77**	**0.78**

Logistic Regression

Logistic Regression served as a baseline model and achieved an accuracy of 75.98% and an AUC-ROC of 0.84. It provided balanced performance on both classes but slightly underpredicted poor health cases

Random Forest

Random Forest improved overall accuracy to 77.52% and increased AUC-ROC to 0.85. It showed stronger ability to detect poor health while ranking key predictors such as BMI, income, employment status, and mental health days as most important.

CatBoost

A gradient-boosting model designed for tabular data, delivered the best performance overall, with an accuracy of 78.17% and the highest AUC-ROC at 0.8614. CatBoost performed especially well at learning complex nonlinear relationships and maintained strong recall and F1 scores for predicting poor health.

Interpretation of predictors

Income-related variables such as Income Category, employment, and ability to afford a doctor were among the strongest predictors, indicating that financial stability is closely tied to physical health and contribute to 35.63% in the overall health prediction (CatBoost Model)**.** BMI, Chronic conditions (diabetes, arthritis), disability limitations, and mental health also played major roles. This suggests that health outcomes are shaped not only by medical conditions, but by broader social and economic determinants.

Interpretation

The results suggest that socioeconomic conditions, financial barriers, and disability limitations play central roles in shaping health outcomes. Individuals with lower income, unstable employment, functional impairments, or chronic health conditions were substantially more likely to report poor health. Conversely, individuals with higher income, more stable employment, and fewer disability limitations had better health outcomes.

The superiority of the CatBoost model further indicates that the relationship between financial stressors and health is nonlinear and complex**,** and traditional linear models may underestimate the combined effects of these factors. Overall, the findings highlight the importance of targeting economic vulnerability, mobility limitations, and long-term chronic conditions in public health policy.

Limitations and Considerations

While this study provides meaningful insights into the relationship between income, socioeconomic factors, and health outcomes, several considerations should be noted:

1.Self-Reported Survey Data

BRFSS relies entirely on self-reported information. Variables such as general health status, exercise, chronic conditions, and mental health days may contain bias**,** misreporting**, or** recall errors**.**

2.Geographically Matched Undersampling Approach

The original data set had a strong class imbalance, with only about 17% of respondents reporting poor health. To improve the model's ability to learn meaningful patterns, we used the state-based random sampling method where poor-health cases were matched with good-health cases within each state. This approach preserves geographic distribution and creates a more statistically reliable prediction framework. Although the resulting dataset is balanced (50/50), this was done intentionally to strengthen the model's learning process not to represent the true national prevalence of poor health.

3. Lack of Clinical Medical Records

This study is based solely on survey responses. Combining BRFSS survey data with actual clinical records in future work would improve prediction accuracy and provide clearer medical insights.

Conclusion

This study examined how socioeconomic, behavioral, and health-related factors predict self-reported health outcomes using BRFSS 2024 data. After extensive cleaning and balancing, three machine-learning models were compared. While Logistic Regression offered clear interpretability, and Random Forest improved predictive accuracy, CatBoost provided the strongest overall performance**,** effectively capturing the complex interactions of financial, demographic, and health-related factors.

Across all models, income-related variables especially Income Category, employment, and affordability of medical care , emerged as some of the most influential predictors of poor health. This reinforces longstanding public health findings: financial stress and socioeconomic disadvantage have powerful impacts on health outcomes**.**

These results highlight the importance of addressing economic inequities, improving access to healthcare, and supporting individuals with chronic conditions and disabilities. While limitations prevent us from establishing causality, the models offer meaningful insights into how social determinants shape health and provide a foundation for deeper future analysis.

References

Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System (BRFSS): Overview and Data for 2024."
www.cdc.gov/brfss/annual_data/annual_2024.html

World Health Organization. "Social Determinants of Health." https://www.who.int/health-topics/social-determinants-of-health

National Academies of Sciences, Engineering, and Medicine. Integrating Social Care into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health. Washington, DC: National Academies Press; 2019. https://nap.nationalacademies.org/catalog/25467/integrating-social-care-into-the-delivery-of-health-care