Analyzing NHIS Data to Investigate Healthcare Access, Mental Health, and Well-being in the United States

Lakshmi Sabari Priya, Jaini
New York University, lj2330@nyu.edu
Siddharth, Shah
New York University, ss16912@nyu.edu

This project delves into healthcare access and mental health disparities in the U.S. using the National Health Interview Survey (NHIS) data. With a dataset of 27,651 samples, we leverage visualizations to uncover patterns, including the impact of ethnicity on access barriers and age-wise distribution of mental health issues. Key variables such as ethnicity, education, regions, drinking status, marital status, and mental health indicators are explored. Predictive analytics tasks focus on binary classification of mental health status and healthcare accessibility, utilizing logistic regression, random forest, and support vector machines. Results aim to inform policymakers and practitioners, offering insights for targeted interventions and equitable healthcare access. This project showcases the power of analytics to address pressing public health challenges and advocate for inclusive healthcare services.

Additional Keywords and Phrases: Data Visualization, Predictive Analytics, Logistic Regression, Support Vector Machine, Random Forest, Healthcare

1 INTRODUCTION

Investigating healthcare access, mental health, and well-being in the United States is a highly compelling data science/analytics project. It addresses critical public health issues and leverages the power of data to drive meaningful change. By extracting insights from extensive healthcare and mental health data, this project offers the opportunity to directly impact millions of lives, making healthcare more accessible and mental health support more effective. Furthermore, it provides an opportunity to address disparities and advocate for equitable healthcare services, benefiting disadvantaged communities.

Beyond its societal significance, this project offers personal fulfillment by contributing to the welfare of others. It's an inspiring chance to improve lives and raise public awareness about health issues. This work opens doors to a growing field of healthcare analytics and related careers, making it an intellectually stimulating and professionally rewarding endeavor. Ultimately, the project's dual promise of substantial societal impact and individual growth makes it a compelling choice for our team.

1.1 Objectives and Goals

This project aims to explore two critical objectives using NHIS data:

- Examine the Accessibility, Usage Patterns, and Disparities in Healthcare Services: Analyze the factors affecting healthcare access and utilization and examine disparities in access to care, such as by race, ethnicity, income, education level, etc.

- Analyze Mental health and well-being trends and factors: Assess the prevalence of mental health conditions in the US population and identify sociodemographic factors affecting mental health issues.

1.2 Dataset Description

- The National Health Interview Survey (NHIS) from the Centers for Disease Control and Prevention (CDC) is our primary data source.
- The NHIS was conducted to collect accurate and up-to-date statistical information on the amount, distribution, and effects of illness and disability in the United States, along with the services provided in response to these health conditions. The data collected is used to study various health issues and trends, including the prevalence of chronic conditions, barriers to accessing healthcare, and sociodemographic factors affecting health. The National Center for Health Statistics (NCHS) and the CDC are responsible for collecting this data. The funding for NHIS comes primarily from the US government.
- Among other publications, National Health Statistics Reports provide in-depth analysis of specific health-related topics using data from the NHIS data. They offer comprehensive insights into various health issues, including disease prevalence, healthcare access, and health behaviors.
- The dataset consists of 27,651 samples with 637 variables. The data comes from a random sample of the entire population of the US.

2 METHODOLOGY

2.1 Data Pre-processing

- **Dependencies and Missing Values** - We analyzed the dependencies discovered in the previous phase and dealt with all these columns by understanding what the missing values meant in each column and filled them accordingly. For example, DIBTYPE_A (Diabetes type) is asked only if DIBEV_A (ever had Diabetes) = 1. So, we filled the missing values in DIBTYPE _A with a value -1, which indicates no Diabetes. There were several interdependent variables of this kind, with some having very complex dependencies which related multiple variables.

- **Data Transformation** - After analyzing all the columns, we adjusted some categories of certain columns based on our requirements. This involved merging or grouping together specific categories of the data to better suit our needs. For example, DIBEV_A had categories 7, 8, and 9, which indicated Refused, Not Ascertained, and Don't know, which all basically meant unknown data. Hence, we treated them as missing values and replaced them with the mode of that column. We also combined the 2 mental health-related attributes (anxiety and depression) in the dataset into one column called MH_STATUS, indicating if the person ever had a mental health illness. We also combined the 5 healthcare accessibility attributes (ABINSUR_A, ABAVAIL_A, ABOPEN_A, ABTOOLONG_A, ABTIME_A) in the dataset into one column called HC_ACCESS indicating if the person has good healthcare access for predictive analytics tasks. Next, we normalize the data using the min-max normalization method to ensure that different features or variables were on the same scale and had equal weight in the analysis, thereby improving the overall quality and accuracy of the results.

- **One-hot encoding** - We applied the technique of one-hot encoding to transform our categorical data into numerical data that could be used as input for our predictive analytics models. This involved creating a binary variable for each category within our categorical variables, where the binary variable would take on the value of 1 if the category was present and 0 otherwise. To perform one hot encoding, we used Python libraries such as scikit-learn or pandas.

2.2 **Data Visualization**

This section employs a diverse range of visualizations to unravel critical insights from the extensive National Health Interview Survey (NHIS) dataset, consisting of 27,651 samples. The aim is to shed light on healthcare access and mental health disparities in the U.S., focusing on key variables such as ethnicity, education, regions, drinking status, marital status, and mental health indicators.

2.2.1. Mental Health in the U.S.

- **Education Levels and Mental Health:** A bar graph illuminates the relationship between education levels and the ratio of individuals experiencing depression or anxiety (see figure 1). This visualization provides a nuanced understanding of how educational attainment correlates with mental health outcomes.

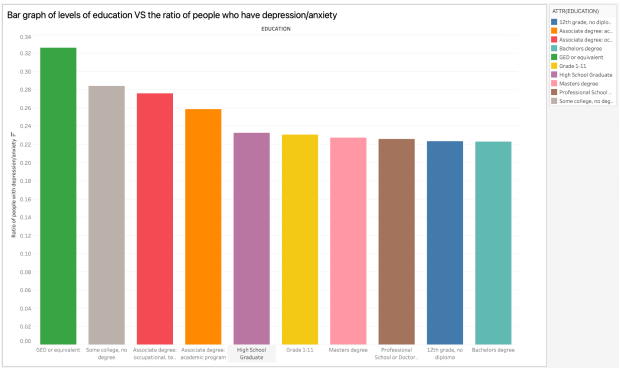

Figure 1: Bar graph of levels of education vs the ratio of people who have depression/anxiety

- **Age, Sex, and Mental Health:** A bar graph further breaks down the ratio of individuals with depression or anxiety by age group and sex (see figure 2). This detailed analysis aims to uncover potential gender-specific or age-specific disparities in mental health.

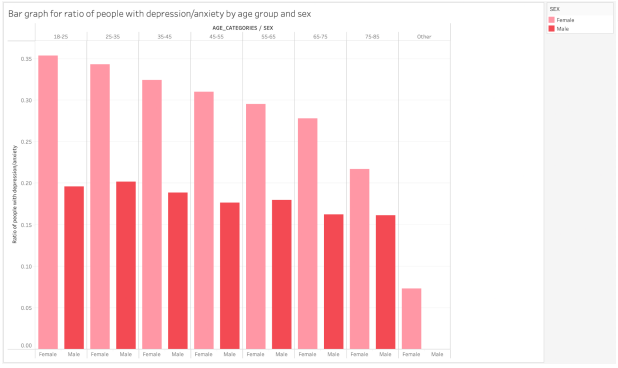

Figure 2: Ratio of people with depression/anxiety by age group and sex

- **Drinking Status Impact on Mental Health:** Ratios of individuals with depression or anxiety are visualized based on drinking status (see figure 3). This provides a comprehensive overview of how lifestyle factors like drinking frequency intersect with mental health indicators.

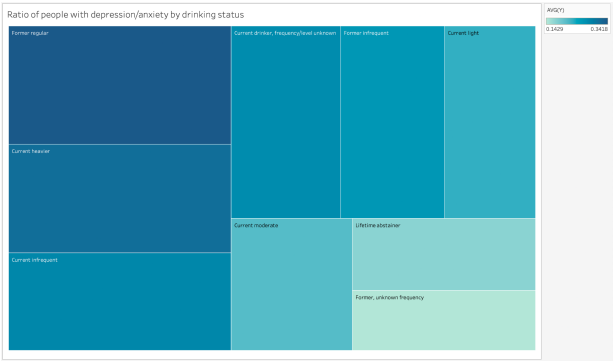

Figure 3: Ratio of people with depression/anxiety by drinking status

- **Marital and Meditation Status Impact on Mental Health:** Ratios of individuals with depression or anxiety are visualized based on their marital and meditation status (see figure 4). This provides an overview of how marriage and lifestyle factors like practicing meditation intersect with mental health indicators.

Ratio of people with depression/anxiety by marital status and meditation status

	MARITAL_STATUS						
MEDITATE_STATUS	Divorced	Living with a partner	Never married	Widowed	Married, spouse is not present	Separated	Married, spouse is present	Married, spouse presence unknown
No	0.2851	0.2551	0.2450	0.2330	0.2330	0.2179	0.1760	0.0000
Yes	0.4055	0.4467	0.4055	0.3219	0.4103	0.4030	0.3211	

Figure 4: Ratio of people with depression/anxiety by marital and meditation status

2.2.2. Healthcare Access in the U.S.

- **Ethnicity and Healthcare Access:** A bar graph delves into the ratio of individuals with good access to care across different ethnicities (see figure 5). This visualization uncovers disparities in healthcare access, emphasizing the importance of addressing ethnic-specific challenges.

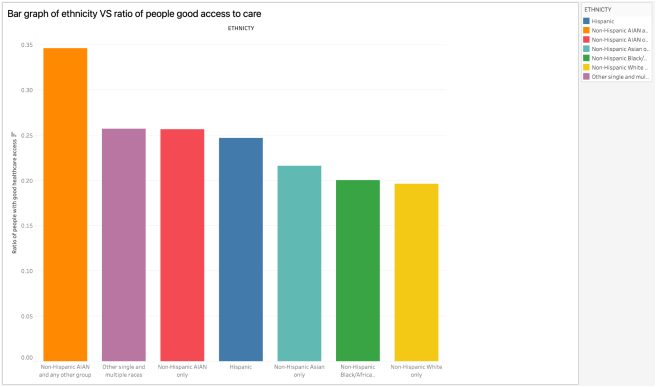

Figure 5: Ethnicity vs ratio of people with good access to care

- **Regional Disparities in Healthcare Access:** Ratios of individuals with good healthcare access are explored across different regions of the USA (see figure 6). This geographical analysis provides insights into regional variations and informs targeted interventions.

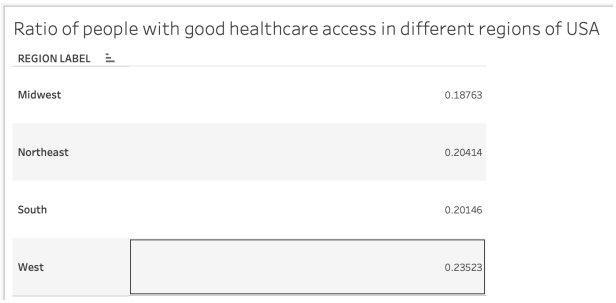

Figure 6: Ratio of people with good healthcare access in different regions of USA

- **Education Levels and Access to Care:** A bar graph investigates the ratio of individuals with good access to care based on their education levels (see figure 7).

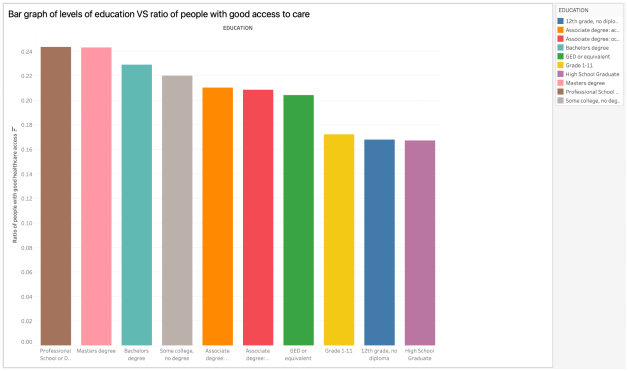

Figure 7: Levels of education VS ratio of people with good access to care

- **Reasons for Lack of Health Insurance:** A pie chart visually represents the reasons why individuals lack health insurance (see figure 8).

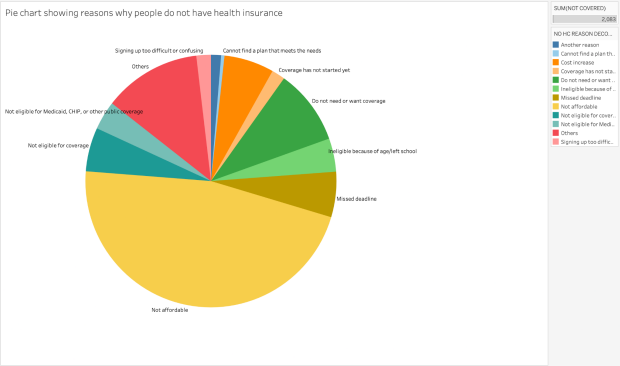

Figure 8: Reasons why people do not have health insurance

These visualizations serve as powerful tools for interpreting complex patterns within the NHIS data, contributing to a comprehensive understanding of mental health and healthcare access disparities. The insights derived from these visualizations are crucial for informing evidence-based policymaking and facilitating targeted interventions to ensure equitable healthcare services.

2.3 **Predictive Analytics**

In this section, we outline the methodology and approach adopted for predictive analytics tasks with the goal of predicting mental health status and associating a healthcare accessibility score based on the diverse set of variables available in the dataset. We employ three popular machine learning algorithms: Logistic Regression, Support Vector Machine (SVM), and Random Forest.

2.3.1. Mental Health Status Prediction:

Model Selection:

- Logistic Regression: Utilize logistic regression for its simplicity and interpretability in binary classification tasks. Assess the impact of various features on mental health status.
- Support Vector Machine (SVM): Employ SVM to handle non-linearity and capture complex relationships within the data. Experiment with different kernels (linear, polynomial, radial basis function) for optimal performance.
- Random Forest: Leverage the ensemble nature of Random Forest to capture intricate patterns in the data, handle non-linearity, and provide feature importance scores.

Model Evaluation:

- Train-Test Split: Divide the dataset into training and testing sets to evaluate model generalization.
- Performance Metrics: Utilize metrics such as accuracy, precision, recall, and F1-score to assess the model's ability to predict mental health status.

2.3.2. Healthcare Accessibility Score Prediction:

Model Selection:

- Logistic Regression: Employ logistic regression to predict a binary healthcare accessibility score (0 or 1). This helps in understanding the factors influencing accessibility.
- Support Vector Machine (SVM): Utilize SVM for regression to predict a continuous score between 0 and 1, representing healthcare accessibility.
- Random Forest: Adapt Random Forest for regression to predict the healthcare accessibility score.

Model Evaluation:

- Train-Test Split: Divide the dataset into training and testing sets to evaluate model generalization.
- Evaluation Metrics: Assess model performance using metrics like accuracy, precision, recall, F1-score.

This predictive analytics section outlines the steps for predicting mental health status and healthcare accessibility scores using Logistic Regression, SVM, and Random Forest. The binary classification approach provides insights into mental health and healthcare accessibility, aiding decision-making and resource allocation. The chosen models balance interpretability and predictive power for actionable results.

3 RESULTS

3.1. Data Visualizations

3.1.1. Mental Health in the U.S.

- **Education Levels and Mental Health:** From this visualization, we can observe that by-and-large a higher level of education leads to lesser mental health problems like anxiety or depression.

- **Age, Sex, and Mental Health:** In this visualization, we can clearly observe that females are more prone to having depression/anxiety than males (in each age group). This inference can be used to target dedicated campaigns for improving womens' mental health.

- **Drinking Status Impact on Mental Health:** From this visualization, it can be observed that current or former heavy drinkers are more prone to mental health issues as compared to lifetime abstainers. This may indicate that people with mental health issues often resort to alcohol in an attempt to escape their mental struggles.

- **Marital and Meditation Status Impact on Mental Health:** Here, it can be observed that people who meditate have higher rates of mental health issues. This may be because most people who struggle with mental health problems resort to some form of meditation as it is supposed to have a good effect on people's mental health. Secondly, divorced people also suffer through depression and anxiety more often than others, as per data.

3.1.2. Healthcare Access in the U.S.

- **Ethnicity and Healthcare Access:** From this visualization, we can see that non-hispanic AIAN and any other group has significantly higher rates of healthcare accessibility than any other ethnic group.

- **Regional Disparities in Healthcare Access:** Here, it can be observed that the West zone in U.S.A. has the highest rate of healthcare accessibility.

- **Education Levels and Access to Care:** It can be observed from this visualization that the more is the level of education of a person, the higher are the chances of them having better healthcare accessibility. This is because Professional school, Doctoral degree, Masters degree, and Bachelors degree have the highest rates of healthcare access.

- **Reasons for Lack of Health Insurance:** Here, it can be seen that almost a majority of the people without an insurance coverage state that the insurance is very expensive and outside the budget.

3.2. Mental Health Status Prediction

In this section, we present the outcomes of our predictive models for mental health status using Logistic Regression, Support Vector Machine (SVM), and Random Forest algorithms. The primary goal is to evaluate the models' effectiveness in distinguishing between individuals with good and bad mental health.

- **Logistic Regression:** The Logistic Regression model achieved an overall accuracy of 80%. In identifying individuals with good mental health (class 0), the model demonstrated high precision (81%) and recall (95%). However, its performance was less satisfactory for individuals with bad mental health (class 1), where precision and recall were 70% and 33%, respectively.

- **Support Vector Machine (SVM):** The SVM model exhibited an accuracy of 77%, displaying notable precision (78%) and recall (98%) for individuals with good mental health (class 0). Nevertheless, the model faced challenges in correctly identifying individuals with bad mental health (class 1), resulting in lower precision (70%) and recall (16%).

- **Random Forest:** The Random Forest model achieved an accuracy of 79%. It excelled in identifying individuals with good mental health (class 0) with precision and recall values of 80% and 97%, respectively. Conversely, the model demonstrated limitations in recognizing individuals with bad mental health (class 1), with precision and recall at 71% and 25%, respectively.

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	Precision (Class 1)	Recall (Class 1)
Logistic Regression	80%	81%	95%	70%	33%
SVM	77%	78%	98%	70%	16%
Random Forest	79%	80%	97%	71%	25%

3.3. Healthcare Accessibility Score Prediction

This section outlines the findings of our healthcare accessibility predictive models, employing Logistic Regression, Support Vector Machine (SVM), and Random Forest algorithms.

- **Logistic Regression:** The Logistic Regression model yielded an overall accuracy of 80%. In distinguishing individuals with good healthcare accessibility (class 0), the model demonstrated commendable precision (82%) and recall (97%). However, its performance for individuals with poor healthcare accessibility (class 1) showed room for improvement, with precision and recall at 63% and 20%, respectively.

- **Support Vector Machine (SVM):** The SVM model achieved an accuracy of 79%, displaying notable precision (80%) and recall (98%) for individuals with good healthcare accessibility (class 0). Nevertheless, challenges emerged in accurately identifying individuals with poor healthcare accessibility (class 1), resulting in diminished precision (56%) and recall (11%).

- **Random Forest:** The Random Forest model secured an accuracy of 80%. It excelled in identifying individuals with good healthcare accessibility (class 0), boasting precision and recall values of 81% and 98%, respectively. Conversely, limitations surfaced in recognizing individuals with poor healthcare accessibility (class 1), with precision and recall standing at 63% and 13%, respectively.

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	Precision (Class 1)	Recall (Class 1)
Logistic Regression	80%	82%	97%	63%	20%
SVM	79%	80%	98%	56%	11%
Random Forest	80%	81%	98%	63%	13%

4 DISCUSSION

- **Mental Health Prediction:** The evaluation of our predictive models suggests that while Logistic Regression and Random Forest show promise in distinguishing mental health status, the SVM model, despite its high accuracy, faces challenges in capturing individuals with bad mental health. Further refinement and exploration of additional features are warranted to enhance the models' predictive capabilities. The results underscore the complexities of predicting mental health status, highlighting the need for continuous model refinement and consideration of diverse factors influencing mental well-being. The next steps in our research involve feature engineering, hyperparameter tuning, and potentially incorporating external factors to improve model robustness and generalizability.

- **Healthcare Accessibility Score Prediction:** These findings highlight the nuanced nature of healthcare accessibility prediction. While Logistic Regression and Random Forest exhibited promise, the SVM model encountered challenges in identifying individuals with poor healthcare accessibility. Further refinement, including feature engineering and model tuning, is crucial to enhance predictive capabilities. These results emphasize the intricacies involved in healthcare accessibility prediction, underlining the necessity for ongoing model refinement and a comprehensive understanding of diverse factors influencing accessibility. Future steps in our research involve meticulous feature engineering, hyperparameter tuning, and potential integration of external variables to fortify model robustness and generalizability.

5 CONCLUSION

This project, delving into healthcare access and mental health disparities in the U.S. through the lens of the National Health Interview Survey (NHIS) data, has successfully harnessed analytics to illuminate critical patterns and challenges. Through visually compelling representations, we unraveled nuanced connections between education, gender, drinking habits, and meditation practices with mental health outcomes. Simultaneously, our exploration of healthcare accessibility unveiled disparities across ethnic groups, regions, and education levels. Leveraging advanced predictive modeling techniques, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, we achieved commendable accuracy in predicting mental health status and healthcare accessibility. While these models offer valuable tools for targeted

interventions, they also underscore the complexities inherent in understanding mental health nuances and predicting healthcare access.

6 REFERENCES

1. National Health Interview Survey. [Online] Available: https://www.cdc.gov/nchs/nhis/2022nhis.htm
2. Terlizzi, E.P., Schiller, J.S. (2021). Mental Health Treatment Among Adults Aged 18–44: United States, 2019–2021. CDC Data Brief, No. 444. [Online] Available: https://www.cdc.gov/nchs/data/databriefs/db444.pdf
3. Marmot, M., Wilkinson, R.G. (2005). The Social Determinants of Health: It's Time to Consider the Causes of the Causes. Public Health Reports, 120(3), 197-205. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3863696/
4. Choi, S., Park, S., Kim, S. (2021). Barriers to healthcare access among U.S. adults with mental health challenges: A population-based study. PLoS ONE, 16(8), e0255854. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8214217/
5. Alagiakrishnan, K., Lim, D., Brahim, A., Wong, A. (2020). Data Analytics in Mental Healthcare. Scientific Programming, 2020, 2024160. [Online] Available: https://www.hindawi.com/journals/sp/2020/2024160/
6. Nagarajan, P., Chockalingam, V., Rajkumar, R.P. (2020). A Comprehensive Analysis of Mental Health Problems in India and the Role of Mental Asylums. Indian Journal of Psychological Medicine, 42(6), 561-567. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10460242/