Heart Health Indicators

Venkata Lakshmi Prasuna Sai Poojitha Anuvala Setty
Data Analytics Engineering
George Mason University
Fairfax, Virginia
vanuvala@gmu.edu , G01421210

***Abstract*—This research explores the data from the Behavioral Risk Factor Surveillance System (BRFSS), a national health-related telephone survey, to investigate patterns in heart diseases among the residents of the U.S. Few attributes that related to heart diseases like health issues and lifestyle factors are extracted and the relationship between them are studied to figure out the likelihood of heart-related conditions. Applying exploratory data analysis techniques, logistic regression in R, Python for visualizations, and AWS-based SQL, the research yields significant findings. Heart disease is most prevalent among men aged 65 and above, with a higher incidence in white individuals compared to other races. Individuals with diabetes, COPD, and kidney issues face elevated risks, emphasizing the impact of health conditions on heart problems. Apart from the health conditions, lifestyle choices play a crucial role in reporting more heart-related issues. Among smokers, the frequency of smoking interacts with physical activity and alcohol habits, influencing heart disease risk. These results emphasize the need for inclusive public health interventions considering all the factors tying to heart diseases. This study offers new insights by using various analytical methods to tackle important questions in heart health.**

***Index Terms*—Behavioral Risk Factor Surveillance System (BRFSS), Exploratory data analysis, Heart Diseases, Logistic Regression**

I. Introduction

Heart diseases pose a significant threat to global public health, leading to high mortality rates and placing a strain on both individuals and healthcare systems. As the primary cause of death worldwide, understanding the complexities of heart diseases is crucial for effective prevention and treatment. The prevalence of these diseases, influenced by factors such as age, gender, lifestyle choices, and underlying health conditions, demands a closer examination. The impact of heart diseases extends beyond statistics, affecting individuals, families, and economies. The economic burden includes healthcare costs, productivity loss, and strain on healthcare infrastructure, while the personal toll encompasses physical, mental, and emotional well-being. This research, utilizing data from the Behavioral Risk Factor Surveillance System (BRFSS), seeks to contribute to a comprehensive understanding of heart diseases. By exploring patterns and relationships, it aims to address key questions about the prevalence and influencing factors of heart-related conditions. The study emphasizes the importance of ongoing research to inform policies and healthcare practices, with the potential to reduce the incidence of heart diseases and improve overall well-being. Through a multidimensional approach considering demographics, health, and lifestyle factors, this research aims to provide insights for targeted interventions in the global effort to combat heart diseases.

II. Literature Review

1. Using Personal Key Indicators and Machine Learning-based Classifiers for the Prediction of Heart Disease Heart diseases are the leading causes of death in the United States. It uses factors like BMI, Smoking, Drinking, Race and so on for analysis. The main purpose of their research is to highlight the importance of integrating machine learning into heart disease prediction and alert people the dangers that are ahead of them using personal indicators. **2. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques** The dataset contains various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC)), and Second-degree block (BII) to find out the exact condition of the patient in relation to heart disease. This paper emphasizes on potential life-saving impact of early detection and preventative measures for heart diseases. It promotes approaches to addressing the complexities of heart disease prediction through machine learning methodologies. **3. Heart Disease Prediction using Exploratory Data Analysis** The dataset includes variables like age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate, and four types of chest pain. The primary focus is on variables directly linked with heart health and the application of algorithms for predictions.

III. Methodology

To start, the dataset is preprocessed using Python, where null values are removed, and certain column values are converted to numerical format. A correlation matrix is then generated to narrow down the analysis. For research inquiries, logistic regression in R is applied, and Python is used to create informative bar charts. The dataset is efficiently managed by loading it into an S3 bucket, facilitated by AWS. MySQL is employed for executing queries, ensuring a structured exploration of the dataset. This streamlined approach, combining Python, R, and AWS-based MySQL, enables a comprehensive investigation into patterns and relationships related to heart diseases.

TABLE I: NOIR Data Types

Column Name	Description	Data Type
State	State in which the respondent resides	Object
Sex	0 = female 1 = male	Ordinal
GeneralHealth	Would you say that in general your health is on a scale of 1-5: 1 = poor 2 = fair 3 = good 4 = very good 5 = excellent	Nominal
PhysicalHealthDays	Physical illness or injury days in the past 30 days	Nominal
MentalHealthDays	Poor mental health days in the past 30 days	Nominal
LastCheckupTime	About how long has it been since you last visited a doctor for a routine checkup?	Nominal
PhysicalActivities	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? 0 = No 1 = Yes	Nominal
SleepHours	On average, how many hours of sleep do you get in a 24-hour period?	Interval
HadHeartAttack	(Ever told) you had a heart attack, also called a myocardial infarction? 0 = No 1 = Yes	Ordinal
HadAngina	(Ever told) (you had) angina or coronary heart disease? 0 = No 1 = Yes	Ordinal
HadStroke	(Ever told) (you had) a stroke. 0 = No 1 = Yes	Ordinal
HadAsthma	(Ever told) (you had) asthma? 0 = No 1 = Yes	Ordinal
HadSkinCancer	(Ever told) (you had) skin cancer that is not melanoma? 0 = No 1 = Yes	Ordinal
HadCOPD	(Ever told) (you had) C.O.P.D. (chronic obstructive pulmonary disease), emphysema or chronic bronchitis? 0 = No 1 = Yes	Ordinal
HadDepression	(Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)? 0 = No 1 = Yes	Ordinal
HadKidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? 0 = No 1 = Yes	Ordinal
HadDiabetes	(Ever told) (you had) diabetes? 0 = No 1 = Yes	Ordinal
SmokerStatus	Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker	Nominal
RaceEthnicityCategory	Five-level race/ethnicity category	Nominal
AgeCategory	Fourteen-level age category	Nominal
HeightInMeters	Reported height in meters	Ratio
WeightInKgs	Reported weight in kilograms	Ratio
BMI	Body Mass Index (BMI)	Ratio
AlcoholDrinkers	Adults who reported having had at least one drink of alcohol in the past 30 days. 0 = No 1 = Yes	Nominal
CovidPos	Has a doctor, nurse, or other health professional ever told you that you tested positive for COVID 19? 0 = No 1 = Yes	Nominal

IV. Analysis

A. Data Cleaning

The dataset contains 40 columns and 445,133 rows. The dataset also contains null values. It also needs to be converted to numerical format to facilitate further analysis. Python was used to clean the dataset and convert the necessary columns into numerical format. In Fig.1, a snapshot of the dataset can be seen.

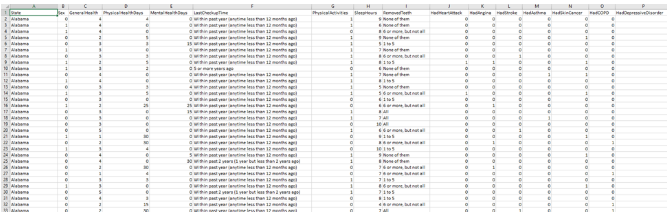


Fig. 1: Boxplots of Factors

B. Univariate Analysis

Univariate analysis involves examining and describing the distribution of a single variable, focusing on measures like mean, median, and dispersion. It provides insights into the characteristics of that variable without considering relationships with other variables, serving as an initial step in statistical analysis. Using AWS Glue DataBrew makes univariate analysis easier.

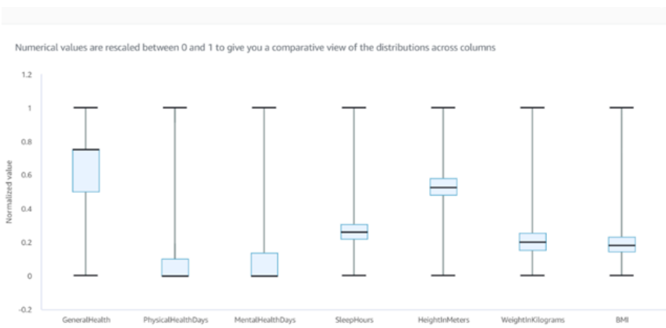


Fig. 2: Boxplots of Factors

The boxplots in Fig.2 reveal that as age increases, the number of heart disease risk factors also rises significantly. Older age groups, especially those between 65-74 and 85+, show higher median risk factors and greater variability. Possible reasons include more chronic health conditions, reduced physical activity, and life stressors in older adults. While the data doesn't prove age causes these risk factors, it aligns with existing research on increased heart disease risk in older individuals. Recommendations for reducing risk include managing chronic conditions, staying active, maintaining a healthy weight, eating well, and avoiding smoking and excessive alcohol. Older adults should discuss their specific risk factors with their doctors.

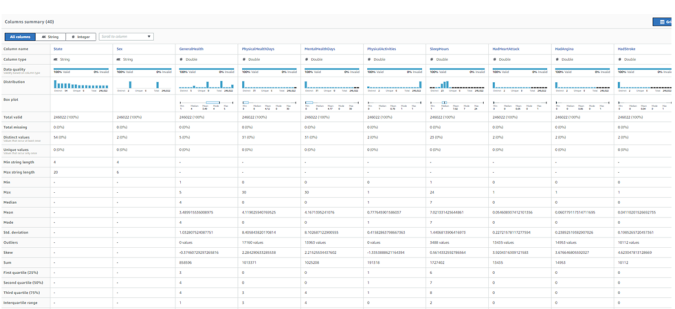


Fig. 3: Summary Statistics of Attributes

This study looked at data from a big survey in the U.S. and found that heart disease is more common in older adults, men, and certain racial/ethnic groups and those with lower income or education. It suggests that efforts to reduce heart disease should focus on these groups and consider social and economic factors like poverty. Overall, the study highlights the need for specific strategies to prevent heart disease based on people's age, gender, and background.

C. Correlation Matrix

The correlation matrix in Fig.5 underscores key associations between variables and heart disease risk factors. Notably, a strong positive correlation exists between the number of health conditions and heart disease risk variables, indicating a higher likelihood of risk factors for individuals with more health conditions. Moderate positive correlations are observed for age and BMI, suggesting older adults and those with higher BMIs are more prone to these risk factors. Conversely, moderate negative correlations are found for physical activity and smoking, indicating that more active individuals and non-smokers are less likely to exhibit these risk factors. While these findings provide valuable insights, it's important to note that correlation does not imply causation. Further research is needed to understand underlying mechanisms and potential confounding factors.

```
selected_variables = df[['PhysicalActivities', 'SmokerStatus', 'AlcoholDrinkers',
                         'HadDiabetes', 'HadHeartAttack', 'HadStroke', 'HadAngina', 'HadAsthma',
                         'HadCOPD', 'HadKidneyDisease', 'BMI', 'CovidPos']]
# Calculate correlation matrix
correlation_matrix = selected_variables.corr()
# Create a heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='flare', fmt=".2f", linewidths=.5)
plt.title('Correlation Heatmap of Health Conditions and Heart Disease Risk Variables')
plt.show()
```

Fig. 4: Code used to generate the Correlation Matrix

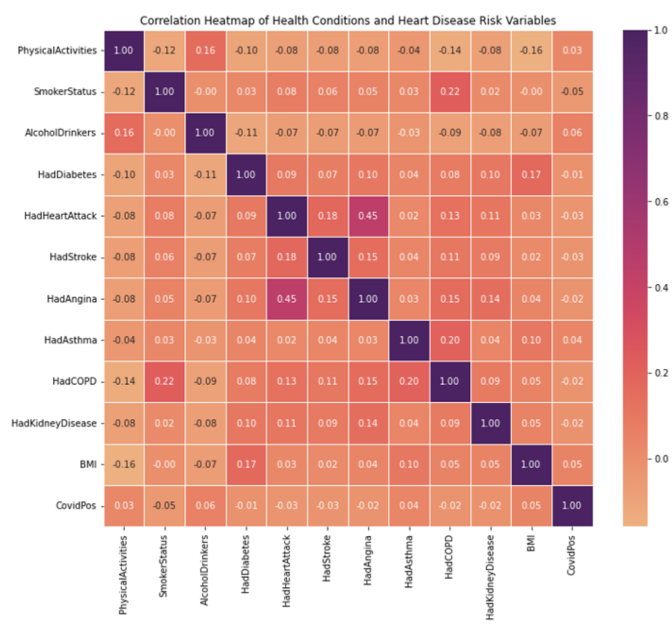


Fig. 5: Correlation Matrix of Indicators

D. Visualization in Python

How do factors like age, sex, and race/ethnicity relate to the likelihood of having a history of heart attack, stroke or angina? To answer this question, a new variable TotalHeart-Diseases is created. It sums the values from three existing columns ('HadHeartAttack', 'HadStroke', 'HadAngina') along each row, providing a consolidated count of heart diseases for each individual in the dataset. Total occurrences of heart diseases are counted against each category. Bar Charts are produced to analyze the statistics.

```python
df['TotalHeartDiseases'] = df[['HadHeartAttack', 'HadStroke', 'HadAngina']].sum(axis=1)

# Combine total occurrences of heart diseases in each AgeCategory
age_category_counts = df.groupby('AgeCategory')['TotalHeartDiseases'].sum().reset_index()

# Combine total occurrences of heart diseases in each RaceEthnicityCategory
race_ethnicity_counts = df.groupby('RaceEthnicityCategory')['TotalHeartDiseases'].sum().reset_index()

# Combine total occurrences of heart diseases in each Sex category
sex_counts = df.groupby('Sex')['TotalHeartDiseases'].sum().reset_index()

# Combine total occurrences of heart diseases in each State
state_counts = df.groupby('State')['TotalHeartDiseases'].sum().reset_index()

# Set up individual subplots
fig, axes = plt.subplots(4, 1, figsize=(15, 25))

# Bar graph for AgeCategory
sns.barplot(x='AgeCategory', y='TotalHeartDiseases', data=age_category_counts, ax=axes[0])
axes[0].set_title('Heart Diseases by Age Category', fontsize=14, fontweight='bold')
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=45, ha='right', fontsize=12)

# Bar graph for RaceEthnicityCategory
sns.barplot(x='RaceEthnicityCategory', y='TotalHeartDiseases', data=race_ethnicity_counts, ax=axes[1])
axes[1].set_title('Heart Diseases by Race/Ethnicity Category', fontsize=14, fontweight='bold')
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=45, ha='right', fontsize=12)

# Bar graph for Sex
sns.barplot(x='Sex', y='TotalHeartDiseases', data=sex_counts, ax=axes[2])
axes[2].set_title('Heart Diseases by Sex', fontsize=14, fontweight='bold')
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=45, ha='right', fontsize=12)

# Bar graph for State
sns.barplot(x='State', y='TotalHeartDiseases', data=state_counts, ax=axes[3])
axes[3].set_title('Heart Diseases by State', fontsize=14, fontweight='bold')
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=45, ha='right', fontsize=12)
```

Fig. 6: Code used for creating bar charts

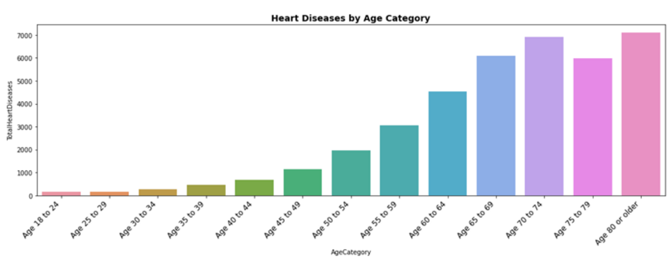


Fig. 7: Heart Diseases by Age

The Fig.7 shows the number of heart diseases in different age categories. It shows that heart diseases are more common among adults in the 65+ category. It also shows that, the occurrence of heart diseases increases with age.

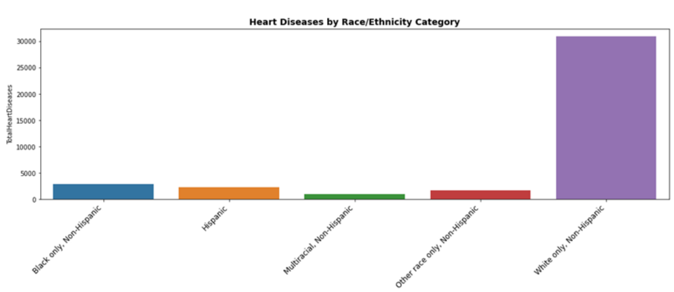


Fig. 8: Heart Diseases by Race/Ethnicity

The Fig.8 shows the number of heart diseases among different race/ethnicity categories. It shows that heart diseases are more common among adults in the White, non-Hispanic category. However, this should be taken into consideration with caution. There could be certain known or unknown biases while conducting the survey or there could be issues with collecting the data.

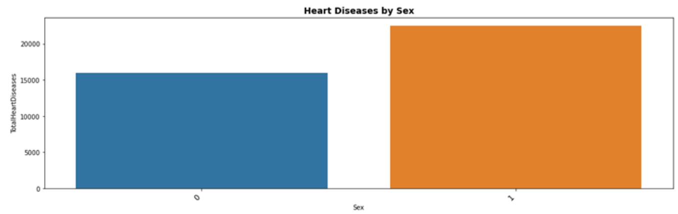


Fig. 9: Heart Diseases by Sex

The Fig.9 shows the number of heart diseases in male and female. It shows that heart diseases are more common among men.

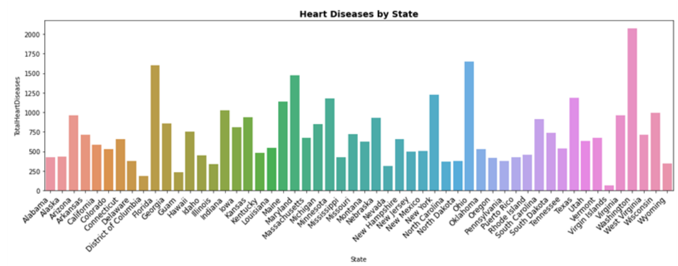


Fig. 10: Heart Diseases by State

In Fig.10, demographic distribution of heart diseases are shown. The states with the highest number of heart diseases are Mississippi, West Virginia, and Arkansas. The states with the lowest number of heart diseases are Utah, Hawaii, and Colorado.

E. Logistic Regression in R

How do people who have multiple health issues like diabetes, high blood pressure, and depression differ from those who don't have these issues when it comes to reporting heart disease events or risk factors? Logistic regression is used in this study to model and understand the complex relationships between health conditions (such as diabetes, COPD, and kidney disease) with the occurrence of heart disease events (heart attack, angina, stroke).

```r
# Create logistic regression models for each heart disease event
model_heart_attack <- glm(HadHeartAttack ~ HadDiabetes + HadDepressiveDisorder + HadCOPD + HadKidneyDisease, data = df, family = "binomial")
model_angina <- glm(HadAngina ~ HadDiabetes + HadDepressiveDisorder + HadCOPD + HadKidneyDisease, data = df, family = "binomial")
model_stroke <- glm(HadStroke ~ HadDiabetes + HadDepressiveDisorder + HadCOPD, + HadKidneyDisease, data = df, family = "binomial")

# Summarize the models
summary(model_heart_attack)
summary(model_angina)
summary(model_stroke)


# Function to calculate accuracy
calculate_accuracy <- function(model, data) {
  # Predict the probability of the event
  predicted_prob <- predict(model, newdata = data, type = "response")

  # Convert probabilities to binary predictions (0 or 1)
  predicted_labels <- ifelse(predicted_prob > 0.5, 1, 0)

  # Create a confusion matrix
  conf_matrix <- table(predicted_labels, data$HadHeartAttack)  # Adjust for each outcome variable

  # Calculate accuracy
  accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
  return(accuracy)
}

# Test the accuracy for each model
accuracy_heart_attack <- calculate_accuracy(model_heart_attack, df)
accuracy_angina <- calculate_accuracy(model_angina, df)
accuracy_stroke <- calculate_accuracy(model_stroke, df)

# Print the accuracies
cat("Accuracy for Heart Attack model:", accuracy_heart_attack, "\n")
cat("Accuracy for Angina model:", accuracy_angina, "\n")
cat("Accuracy for Stroke model:", accuracy_stroke, "\n")
```

Fig. 11: Code used to create regression models for heart diseases

```
> summary(model_heart_attack)

Call:
glm(formula = HadHeartAttack ~ HadDiabetes + HadDepressiveDisorder +
    HadCOPD + HadKidneyDisease, family = "binomial", data = df)

Coefficients:
                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)           -3.23344    0.01199 -269.578   <2e-16 ***
HadDiabetes            0.42026    0.01174   35.784   <2e-16 ***
HadDepressiveDisorder  0.01336    0.02159    0.619    0.536
HadCOPD                1.20940    0.02313   52.280   <2e-16 ***
HadKidneyDisease       1.09962    0.02799   39.283   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104249  on 246021  degrees of freedom
Residual deviance:  98548  on 246017  degrees of freedom
AIC: 98558

Number of Fisher Scoring iterations: 6
```

Fig. 12: Summary of Heart Attack Model

According to the output in Fig.12,individuals with diabetes, COPD, and kidney disease are significantly more likely to experience a heart attack, with a high model accuracy of 94.53 percent.

```
> summary(model_angina)

Call:
glm(formula = HadAngina ~ HadDiabetes + HadDepressiveDisorder +
    HadCOPD + HadKidneyDisease, family = "binomial", data = df)

Coefficients:
                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)           -3.16776    0.01160 -273.026   <2e-16 ***
HadDiabetes            0.41886    0.01137   36.829   <2e-16 ***
HadDepressiveDisorder  0.02150    0.02061    1.043    0.297
HadCOPD                1.30564    0.02191   59.584   <2e-16 ***
HadKidneyDisease       1.35270    0.02566   52.709   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112730  on 246021  degrees of freedom
Residual deviance: 104881  on 246017  degrees of freedom
AIC: 104891

Number of Fisher Scoring iterations: 5
```

Fig. 13: Summary of Angina Model

According to the output in Fig.13, diabetes, COPD, and kidney disease as highly significant predictors (p ¡ 2e-16 for all), indicating a strong association with the likelihood of experiencing angina. The model demonstrated a high accuracy of 94.52 percent, emphasizing its effectiveness in predicting angina events.

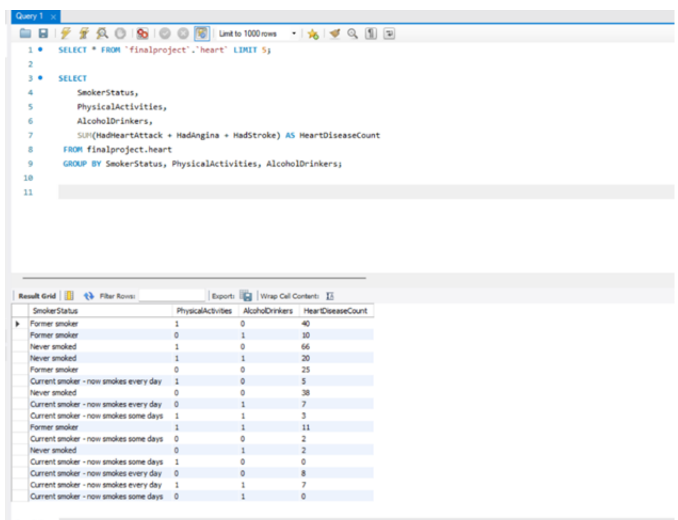
```
> summary(model_stroke)

Call:
glm(formula = HadStroke ~ HadDiabetes + HadDepressiveDisorder +
    HadCOPD, family = "binomial", data = df, weights = +HadKidneyDisease)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.35838    0.04501 -52.392  < 2e-16 ***
HadDiabetes            0.28323    0.04075   6.951 3.62e-12 ***
HadDepressiveDisorder  0.26689    0.06097   4.378 1.20e-05 ***
HadCOPD                0.69532    0.06536  10.639  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8446.3  on 11283  degrees of freedom
Residual deviance: 8248.6  on 11280  degrees of freedom
AIC: 8256.6

Number of Fisher Scoring iterations: 4
```

Fig. 14: Summary of Stroke Model

According to the output in Fig.14, diabetes, COPD, and depressive disorder as significant predictors (p = 3.62e-12, p ¡ 2e-16, and p = 1.20e-05, respectively), indicating a substantial impact on the likelihood of experiencing a stroke. The model demonstrated a high accuracy of 94.54 percent, highlighting its effectiveness in predicting stroke events.

F. Using MySQL through AWS

The research investigates the correlation between lifestyle factors (SmokerStatus, AlcoholDrinkers, PhysicalActivities) and the likelihood of reporting heart disease-related conditions. This analysis is crucial for informing public health initiatives, aiding risk assessment for individuals and healthcare providers, and guiding policy development to reduce heart disease incidence by addressing modifiable risk factors.



```
SELECT * FROM `finalproject`.`heart` LIMIT 5;

SELECT
    SmokerStatus,
    PhysicalActivities,
    AlcoholDrinkers,
    SUM(HadHeartAttack + HadAngina + HadStroke) AS HeartDiseaseCount
 FROM finalproject.heart
 GROUP BY SmokerStatus, PhysicalActivities, AlcoholDrinkers;
```

SmokerStatus	PhysicalActivities	AlcoholDrinkers	HeartDiseaseCount
Former smoker	1	0	40
Former smoker	0	1	10
Never smoked	1	0	66
Never smoked	1	1	20
Former smoker	0	0	25
Current smoker - now smokes every day	1	0	5
Never smoked	0	0	38
Current smoker - now smokes every day	0	1	7
Current smoker - now smokes some days	1	1	3
Former smoker	1	1	11
Current smoker - now smokes some days	0	0	2
Never smoked	0	1	2
Current smoker - now smokes some days	1	0	0
Current smoker - now smokes every day	0	0	8
Current smoker - now smokes every day	1	1	7
Current smoker - now smokes some days	0	1	0

Fig. 15: Summary of Stroke Model

The data shows that people who used to smoke or never smoked but don't drink alcohol are more likely to report heart disease-related issues. Among current smokers, those who smoke every day and also drink alcohol have a higher likelihood of reporting such conditions. On the other hand, people who smoke some days, especially if they don't drink alcohol, seem to have a lower incidence of heart disease-related problems. These findings suggest that smoking habits, alcohol consumption, and frequency of smoking interact in different ways with heart health. It's essential to consider these factors together to better understand the risk of heart disease-related conditions.

V. Limitations

Analyzing the BRFSS health dataset has some challenges. People might not always give accurate information about their health, and the way they answer survey questions could be influenced by what they think is socially acceptable. The survey is done over the phone, which might leave out some groups of people. For example, while analyzing the presence of heart diseases with respect to race/ethnicity of people, the data shows that the most effected race is White, Non-Hispanic group. This cannot be confirmed without further analysis. The data is like a snapshot, showing us a picture of health at one point in time, but it doesn't tell us how things change over time. We also don't have all the details about people's health, and some important information might be missing. So, while the data is helpful, we need to be careful not to draw conclusions that might not be entirely accurate, and researchers need to consider these limitations when studying the information.

VI. Future Scope

The future scope of analyzing the heart health indicators dataset lies in its potential to advance public health research and policy making. Longitudinal studies can uncover trends over time, predictive modeling can forecast health issues, and precision medicine may benefit from integrating genetic and clinical data with behavioral information. The analysis could deepen our understanding of health disparities, evaluate the effectiveness of interventions, and explore the impact of environmental factors on health outcomes. Integrating heart health indicators data with electronic health records could provide a more comprehensive health profile, while global comparisons may reveal international best practices. As mental health gains attention, the dataset could offer insights into factors affecting mental well-being. Embracing emerging technologies, refining data visualization, and applying machine learning can further enhance the dataset's utility in addressing evolving health challenges. Overall, the future holds exciting possibilities for leveraging heart health indicators data to inform evidence-based strategies for improving public health on a broad scale.

VII. Conclusion

In short, the analysis shows that as people get older, the risk factors for heart disease go up. Older age groups, especially those between 65-74 and 85+, have more risk factors. This suggests that focusing on age-related factors like chronic health conditions and less physical activity could help prevent heart issues. The study also finds that heart disease is more common in older adults, men, and certain racial/ethnic groups, pointing to the need for specific prevention efforts. Looking at the relationships between different factors, the analysis confirms that having more health problems and being older or having a higher BMI increases the likelihood of heart issues. On the other hand, being physically active and not smoking are linked to a lower risk. The study also considers where people live, showing that some states have higher heart disease rates than others. Analyzing various predictors, like diabetes and smoking, indicates their significance in predicting heart disease events. The data also reveals that the combination of smoking, drinking, and how often someone smokes plays a role in heart health. In a nutshell, the research provides insights for tailoring strategies to prevent heart disease, considering factors like age, lifestyle, and where people live.

References

[1] "BRFSS Survey Data 2022", August, 2023 Available: https://www.cdc.gov/brfss/annualdata/annual2022.html

[2] R. Jahed, O. Aseer and A. Al-Mousa, "Using Personal Key Indicators and Machine Learning-based Classifiers for the Prediction of Heart Disease," 2023 International Conference on Smart Computing and Application (ICSCA), Hail, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICSCA57840.2023.10087430.

[3] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[4] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," Procedia Computer Science, vol. 173. Elsevier BV, pp. 130–139, 2020. doi: 10.1016/j.procs.2020.06.017.