Project Proposal on
Analysis of Kaggle's 2015's
Behavioral Risk Factor Surveillance System (BRFSS) dataset

Submitted By

Alina Shrestha (C0901127)
Manjul Silwal (C0899871)
Nirajan Khadka (C0900308)
Pralush Shrestha (C0900891)
Sabina Darlami (C0900811)
Sabina Thapa (C0899899)
Shweta Bhattarai (C0901156)
Utkrista Chapagain (C0901198)

Submitted to
Mr. Vasil Khachidze
Course Instructor
BAM: 1024- Introduction to Statistical Analysis

Submitted as part of course requirement

July 23, 2023

Table of Contents

1. Introduction

The Kaggle's 2015’s *Behavioral Risk Factor Surveillance System (BRFSS)* dataset is a comprehensive health telephone survey conducted by the Centers for Disease Control and Prevention (CDC) to track various behavioral risk factors, chronic health conditions, and health prevention strategies. This project aims to gain valuable insight into the behavioral and health-related factors that influence the prevalence of heart disease and explore the possibilities of predictive modeling to aid early detection and intervention.

This project aims to identify patterns and trends related to heart disease prevalence from this dataset. Using advanced machine learning techniques, a predictive model will be developed that can evaluate an individual's risk of developing heart disease. This model can help experts in evaluations of dangers, personal interventions and care for persons and health collectors.

2. About the Data

The dataset contains real data collected from various individuals in the United States, making it highly relevant for public health research. It contains a wide range of health indicators, including demographic information such as age, gender, and education level, as well as behavioral risk factors such as smoking, physical activity and dietary habits. In addition, the dataset contains important physiological measurements such as blood pressure, cholesterol level and body mass index (BMI). The richness and scope of this dataset provides a unique opportunity to study the complex interaction of behavioral and health factors associated with heart disease prevalence.

The dataset is a comprehensive collection of health-related which includes the same differences and continuous variables, making it suitable for different analytical methods. The dataset captures the diversity of the U.S. population, providing a representative sample allowing generalizations to be made about the broader population. The dataset has a high degree of authenticity and reliability as it is collected by a genuine organization. However, the dataset has issues with certain label of accuracy like class imbalance in this dataset. One of the main concern with this dataset is missing data that arose from respondents who choose not to provide their certain health related information. Additionally, due to different survey approaches, there may be variations in data quality which may require data pre-processing and cleaning to ensure the validity of the analysis.

Datasheet Link is https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv

3. Interesting Questions

This project aims to answer three questions based on the selected subset of the dataset, allowing the project to focus on specific variables that are most relevant to heart disease occurrence, facilitating a deeper and more targeted examination of the dataset.

By focusing on specific health-related variables, valuable opinions regarding the spread of heart disease and the factors concerned, the answers related to the questions will be gained and findings from these questions will contribute to the evidence-based heart disease prevention and intervention strategies, ultimately improving public health outcomes.

a. **How do behavioral risk factors such as the condition of smoking, physical activity and eating habits related to heart disease?**
   This analysis aims to check the connection between the heart disease and the behavioral risk factors, helping promote healthier lifestyle and reduce risk of heart disease.

b. **Are there significant differences in the prevalence of heart disease based on demographic factors such as age, gender, and education level?**
   This analysis aims to investigate potential differences in heart disease occurrence between different population groups, helping design equitable and inclusive public health initiatives.

c. **How well can the prevalence of heart disease be predicted using a combination of behavioral risk factors and physiological measures (example: blood pressure, cholesterol levels, and BMI)?**
   This analysis aims to create a predictive model that includes behavioral risk factors and physiological measures, helping determine person's risk of developing heart disease.

4. Key Variables

During the analysis of this project, the primary target variable representing the “presence” or “absence” of heart disease is a binary categorical indicator i.e. 1 for presence of heart disease and 0 for absence of heart disease. The binary classification framework will serve as the base for our predictive modeling, allowing to distinguish between individuals with or without heart disease. Furthermore, for independent variables, certain subset features that include a combination of behavioral and health indicators from the dataset has been selected based on their relevance with risk factors for heart disease, which are mentioned below:

a. High blood pressure (indicated by _RFHYPE5): Respondents who were told by a doctor that they had high blood pressure.

b. High cholesterol (indicated by TOLDHI2 and _CHOLCHK): Respondents who were told by a doctor that their blood cholesterol level was high and information about their cholesterol control over the past five years.

c. Body Mass Index (indicated by _BMI5): Measurement of respondents' BMI, which is a continuous variable used to assess body weight in relation to height.

d. Smoking (indicated by SMOKE100): Respondents who smoked at least 100 cigarettes in their lifetime.

e. Other health conditions (indicated by CVDSTRK3): The information is when respondents was told by a health care professional that they had.

f. Physical Activity (indicated by _ TOTINDA): Respondents who said they were doing physical activity or practicing the last 30 days, except their usual work.

g. Diet (indicated by _FRTLT1 and _VEGLT1 respectively): Information on whether respondents eat fruits at least once a day and information on whether respondents eat vegetables at least once a day.

h. Alcohol consumption (indicated by _RFDRHV5): Identification of heavy drinkers based on weekly alcohol consumption limits for adult men and women (more than 14 for men and more than 7 for women).

i. Health care (indicated HLTHPLN1 and MEDCOST respectively): Information on whether respondents had any type of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service and whether they encountered barriers related to the cost of visiting a doctor during the past 12 months.

j. General health and mental health (indicated by GENHLTH, MENTHLTH, PHYSHLTH and DIFFWALK respectively): Self-reported overall health status. Number of days in the past 30 days when respondents had poor mental health and physical health (which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?). Information on whether respondents have severe difficulty walking or climbing stairs.

k. Demographics (indicated by _SEX, _AGEG5YR, EDUCA and INCOME2): Demographic variables such as gender, and age group (fourteen-level category) along with information on highest year or year of school completed and annual household income (if respondent refused at any income level, coded "Refused").

For ensuring the accuracy and consistency of the data, reference to the "BRFSS 2015 Code Book" will be taken for a clear understanding of the questions and variables in the dataset. Appropriate data pre-processing and cleaning procedures will be done in the dataset to address these issues and ensure the validity of the analysis. In addition, considering the potential imbalance in the prevalence of heart disease due to its relatively rare occurrence, appropriate methods will be used to reduce bias and improve the accuracy of the predictive model for cases of heart disease and other diseases.

Furthermore, inspiration from related research work by Zidian Xie et al will be taken, who used the same features from the 2014 BRFSS to create risk prediction models for type 2 diabetes. Given the strong correlation between diabetes and heart disease outcomes, this selection of features provides a useful starting point for the analysis.

5. Usefulness of the project

The purpose and benefits of this project are mentioned below:

a. **Heart Disease Prevention and Intervention:** The report generated from this project will provide the information regarding relationship between behavioral risk factors such as smoking, physical activity, and dietary habits and heart disease occurrence.

b. **Equitable Public Health Initiatives:** The report generated from this project will provide the information regarding occurrence of heart disease based on demographic factors, such as age, gender, and education level. This information will be crucial for developing inclusive and equitable public health initiatives that address the specific needs of different population groups.

c. **Predictive Model for Heart Disease Risk:** Developing a predictive model using behavioral risk factors and physiological measures, such as blood pressure, cholesterol levels, and BMI, will enable early identification of individuals at high risk of developing heart disease. This model can assist healthcare professionals in providing timely interventions and personalized care to reduce the burden of heart disease.

This project will can be used by the following users for their reference:

a. **Public Health Authorities:** The report generated from this project will be useful for Public health authorities, such as the Centers for Disease Control and Prevention (CDC) for creating evidence-based policies and interventions to fight heart disease and promote better heart health at a population level.

b. **Healthcare Providers:** The report generated from predictive model's implementation in this project will be useful for physicians, nurses, and other healthcare providers aiding them in risk assessment and early detection of heart disease in patients, allowing for timely interventions and personalized care plans.

c. **Research Community:** The report generated from this project will be useful for researchers in the fields of public health, epidemiology, and cardiovascular disease as this study can serve as a basis for further investigations and studies, contributing to the growing body of knowledge on heart disease risk factors and prevention.

d. **General Public:** The report generated from this project will be useful for the general public as they can be awarded of heart disease risk factors and the importance of healthy lifestyle choices. Public awareness campaigns can be developed based on the findings of this project to educate individuals about heart disease prevention.

6. Responsibilities

    a. Pralush Shrestha:

Data Cleaning and Preprocessing: Cleaning the dataset, handling missing data, addressing outliers, and ensuring data quality.

b. Sabina Darlami:

Exploratory Data Analysis (EDA): Conducting the initial exploratory data analysis to gain insights into the dataset, identifying patterns, and summarizing key trends related to heart disease prevalence.

c. Utkrista Chapagain:

Data Visualization: Creating informative visualizations to effectively communicate the distribution of variables, relationships, and trends in the data.

d. Nirajan Khadka:

Statistical and Hypothesis Testing: Conducting appropriate statistical tests to validate hypotheses and assess the significance of relationships between variables and heart disease prevalence.

e. Alina Shrestha:

Model Development and Evaluation: Leading the efforts in selecting appropriate machine learning algorithms, building the predictive model, and evaluating different models to achieve optimal performance.

f. Shweta Bhattarai:

Interpretation of Results and Insights: Analyzing the model's predictions, interpreting feature importance, and deriving actionable insights from the analysis to guide recommendations and conclusions.

g. Manjul Silwal:

Validation and Sensitivity Analysis: Ensuring the reliability of the predictive model. Describing any validation techniques used, such as k-fold cross-validation or train-test split. Additionally, performing sensitivity analysis to assess the robustness of the model by varying certain parameters.

h. Sabina Thapa:

Future Works: The future work section suggests potential directions for further research and analysis. It may include exploring additional variables, applying more sophisticated modeling techniques, or investigating the impact of specific interventions.

h. All Team Member:

Reporting and Presentation: Collaboratively preparing the final project report and presentation, summarizing the findings, explaining the data analysis process, and presenting the results in a clear and concise manner.

Through this integrated approach, our team strives to effectively analyze the "BRFSS 2015" Dataset, gaining meaningful insights, and effectively presenting our findings and recommendations, delivering a comprehensive and meaningful data analysis project focused on the prevalence of heart disease and its risk factors.

References

(n.d.). https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv