How Does Lifestyle Influence Heart Disease Risk

Team #7

Bara'a Abu-Muhaisen| Esra'a Alhaj| Talab Yaseen| Tala Tawil|Qusai AL-Ma'ani

Table of contents:

Introduction

Heart disease is one of the leading causes of death in the world today. According to the World Health Organization (WHO), 17.9 million people die each year due to cardiovascular diseases. The three major risk factors for heart disease are smoking, high blood pressure and cholesterol. In the United States alone, 47% of all citizens have at least one of these risk factors. In this project, we aim to predict heart disease based on various symptoms using a dataset called “Key Indicators of Heart Disease” from Kaggle. The dataset has 300,000+ instances and 18 attributes, of which one is the class attribute (heart disease) and the rest are predictive attributes. The predictive attributes include gender, age, race, obesity (high BMI), diabetic condition, physical activity level, general health, mental health, alcohol consumption, smoking, stroke status, walking difficulty, asthma, kidney disease, and skin cancer. The class attribute is binary, indicating whether the respondent has heart disease or not. The objective of this project is to predict the most important factors affecting heart disease and to understand how these factors interact with each other. We will use machine learning algorithms through python to analyze the data and identify patterns that can help us predict heart disease more accurately.

Specific Problem

Heart disease is a major public health concern worldwide. According to the World Health Organization (WHO), 17.9 million people die each year due to cardiovascular diseases. While the three major risk factors for heart disease are well-known (smoking, high blood pressure and cholesterol), there are several other factors that can contribute to the development of heart disease1. In this project, we aim to identify the most important factors affecting heart disease and understand how these factors interact with each other.

Our goal is to increase awareness of the indicators that lead to heart disease and provide valuable insights and recommendations for healthcare providers and policymakers to improve the prevention and management of heart disease**.**

Business Impact

Heart disease is a major public health problem that affects millions of people and costs billions of dollars in healthcare and economic losses. Preventing and treating heart disease requires a comprehensive approach that considers various risk factors and indicators that can influence the development and progression of the disease. In this project, we aim to analyze a large-scale health

survey dataset using python and Tableau to identify patterns and relationships between various indicators and heart disease. By analyzing this dataset, we aim to identify patterns and relationships between various indicators and heart disease. Our goal is to increase awareness of the indicators that lead to heart disease and provide valuable insights and recommendations for healthcare providers and policymakers to improve the prevention and management of heart disease.

Data

Used Dataset https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

- We identified a dataset collected by a number of researchers from the CDC under the name: “Key Indicators of Heart Disease” from Kaggle. The dataset contains data from the Behavioral Risk Factor Surveillance System (BRFSS), which is an annual telephone survey that collects data on the health status of U.S. residents.
- The dataset covers the period from 2015 to 2020 and consists of 401,958 rows and 20 columns as detailed in the data curation document. The dataset size is 24 MB.
- In the first stage, we downloaded the dataset from Kaggle and performed data cleaning and preprocessing steps. We performed exploratory data analysis to understand the distribution and correlation of the variables.
- We split the dataset into training and testing sets with an 80:20 ratio; We used the training set to train our prediction models and the testing set to evaluate their performance.

Data Analysis & Computation

Exploratory Data Analysis

Our dataset focuses on the risk factors and indicators of heart disease.The dataset has 401,958 rows and 20 columns, which are selected from the original dataset of 279 columns. The dependent variable is “HeartDisease”, which indicates whether a respondent has heart disease or not.. The independent variables are various demographic, behavioral, and health-related factors, such as age, gender, race, smoking status, physical activity, BMI, blood pressure, cholesterol, and diabetes, after conducting EDA on the dataset we discovered the following findings:

The table (1) shows the descriptive statistics of four variables from the dataset: BMI, PhysicalHealth, MentalHealth, and SleepTime. BMI is a measure of body fat based on height and weight. PhysicalHealth and MentalHealth are the number of days in the past 30 days that the respondent’s physical or mental health was not good. SleepTime is the average number of hours of sleep per night in the past 30 days. The table displays the count, mean, standard deviation, minimum, maximum, and quartiles of each variable. The table reveals that the average BMI of the respondents is 28.32, which is considered overweight. The average PhysicalHealth and MentalHealth are 3.5 and 4.12 days, respectively, indicating that most respondents have good health status. The average SleepTime is 7.1 hours, which is close to the recommended amount of sleep for adults.

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	301717.000000	301717.000000	301717.000000	301717.000000
mean	28.441970	3.572298	4.121475	7.084559
std	6.468134	8.140656	8.128288	1.467122
min	12.020000	0.000000	0.000000	1.000000
25%	24.030000	0.000000	0.000000	6.000000
50%	27.410000	0.000000	0.000000	7.000000
75%	31.650000	2.000000	4.000000	8.000000
max	94.850000	30.000000	30.000000	24.000000

Table 1: descriptive statistics of the numerical variables

Figure 1 shows the distribution of the target variable HeartDisease in the dataset. Heart Disease is a binary variable that indicates whether a respondent has heart disease or not.). The pie chart shows that the majority of the respondents do not have heart disease with 91.4% , while only a small fraction of them do have heart disease (8.6%.

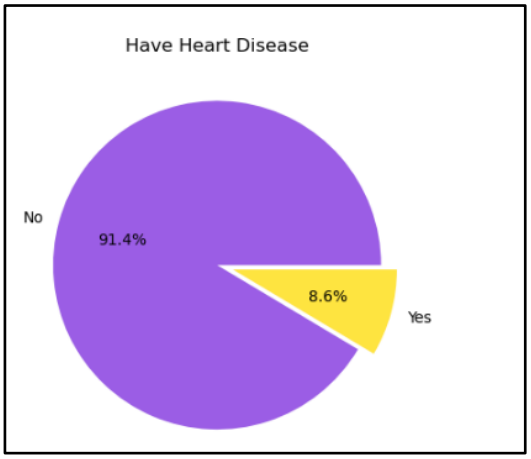

Figure 1: Distribution of the heart disease variable

As shown in Figure (2), the Box plots revealed that four attributes had many outliers: BMI, Physical Health, Mental Health, and Sleep. Subsequently, to detect the outliers, a Mahalanobis Distance Test was performed using SPSS. This test identified 15,654 multivariate outliers out of 301,717 data points. These outliers were removed from the dataset to reduce their influence on the subsequent analysis. The removal of outliers improved the normality and homogeneity of variance assumptions for the data.

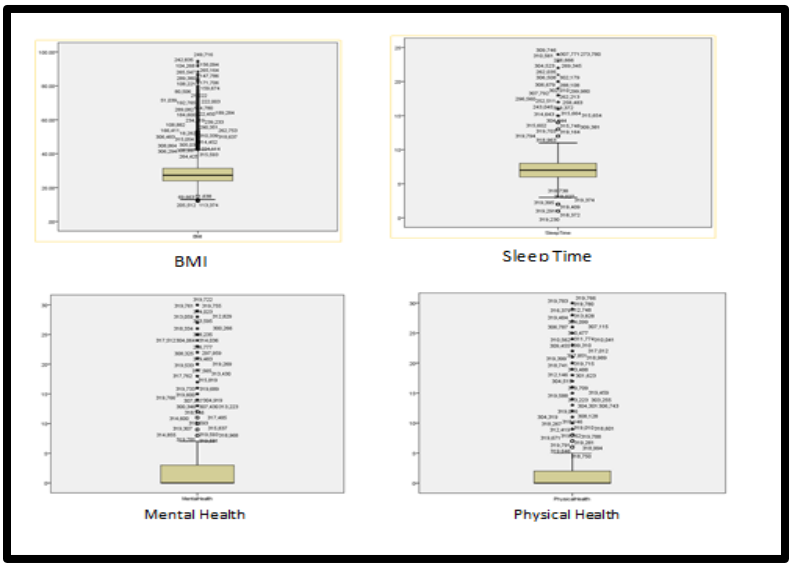

Figure 2 : Box plots for the Numerical variables

The following bar charts show the percentage of heart disease based on three variables: stroke, skin cancer and kidney disease. The bar charts show that there are some differences in the percentage of heart disease based on these variables. The percentage of heart disease is higher for the respondents who have had a stroke (25.9%), have had skin cancer (8.8%), or have had kidney disease (15.5%) than for those who have not. These results suggest that stroke, skin cancer, and kidney disease are positive predictors of heart disease in this dataset.

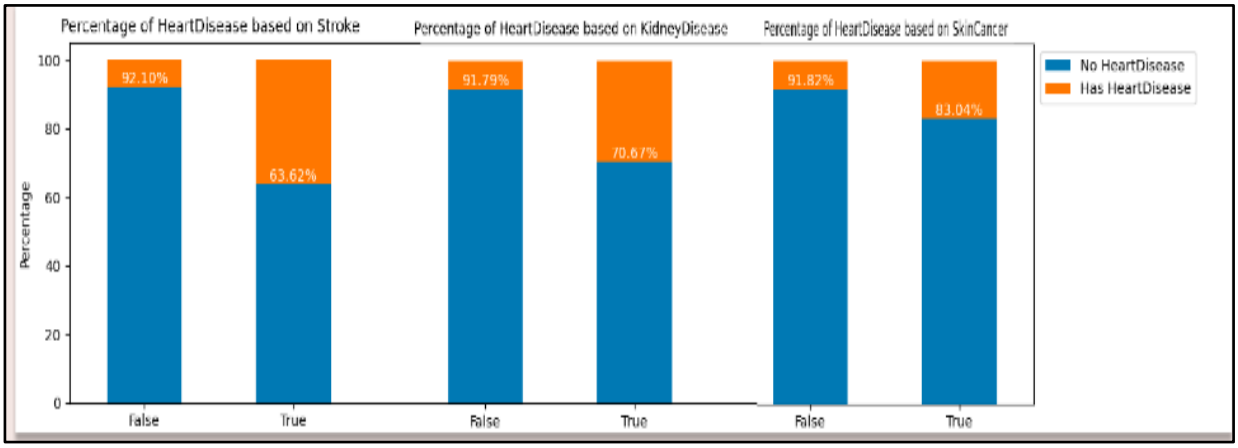

Figure 3: the percentage of heart disease based on three variables: stroke, skin cancer and kidney disease.

The bar charts in figure 4 show that the percentage of heart disease is lower for the respondents who drink alcohol (6.8%) than for those who do not. This suggests that alcohol are negative predictors of heart disease in this dataset. However, this does not mean alcohol is good for the heart, as they can cause other health problems. Smoking and alcohol may be associated with other factors that influence heart disease, such as age, gender, race, or lifestyle.

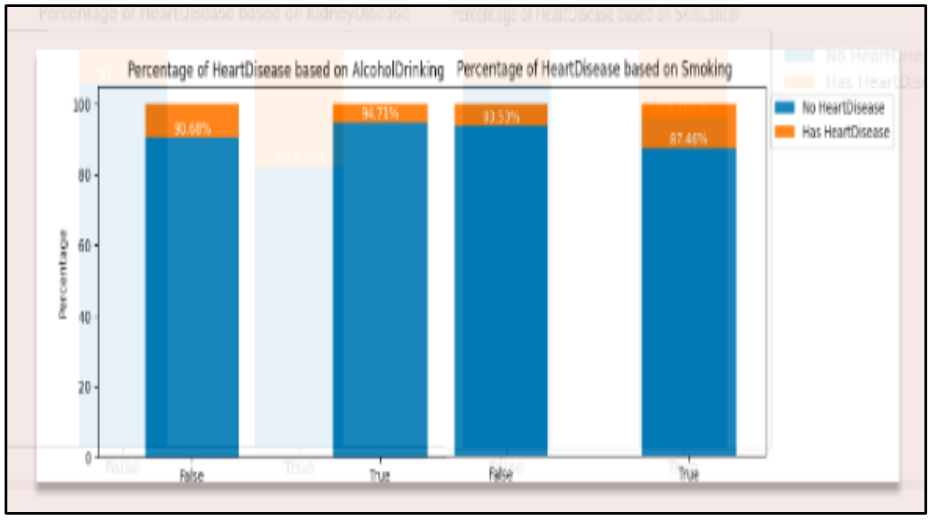

Figure 4 : the percentage of heart disease based on smoking and Alcohol drinking.

The bar chart shows that the percentage of heart disease increases with age. The percentage of heart disease is lowest for the youngest age category (18-38) with 1.8%, and highest for the oldest age category (81-99) with 17.9%. This suggests that age is a positive predictor of heart disease in this dataset.

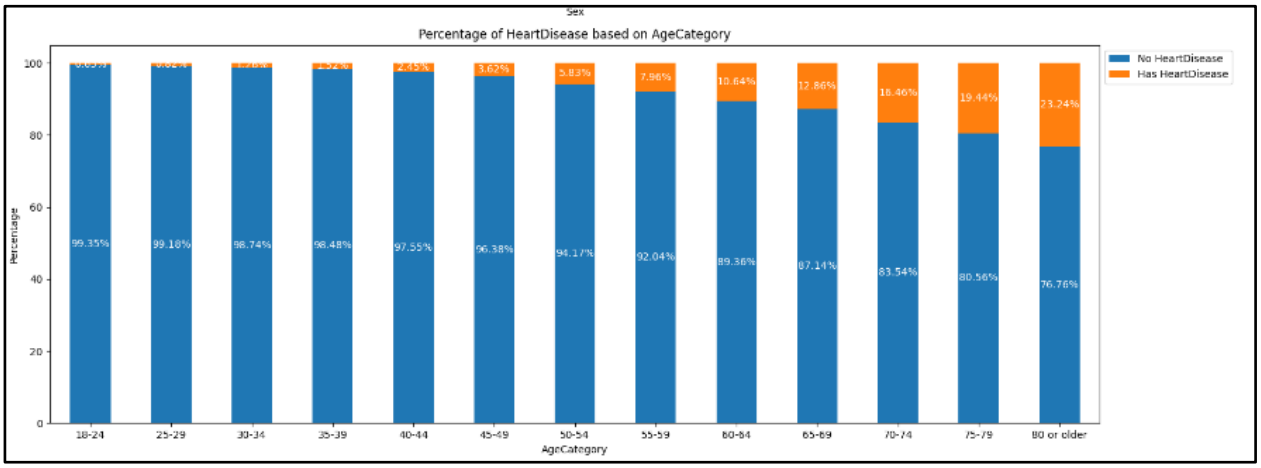

Figure 5: relationship between heart disease and age

Data Wrangling:

- After removing duplicates using Python, the number of data points was reduced to 301,717.
- We explored the data using descriptive statistics such as (mean, standard deviation, min, max, 1/2/3 quartiles)
- We used Python (pandas) to investigate whether the columns had null values or not. The result shows that there were no missing values to deal with.
- We checked for outliers and found that there are some extreme values in the numerical columns such as “Age”, “BMI”, “GeneralHealth”, and “MentalHealth”. We decided to use the Mahalanobis Distance Test to identify and remove the outliers.
- We checked for categorical variables and found that there are 15 categorical variables in the dataset. We decided to use label encoding to convert them into numerical values.
- We performed exploratory data analysis to understand the distribution and correlation of the variables. We used histograms, box plots, scatter plots, and heat maps to visualize the data.
- We conducted train and test and SMOTE techniques because the data was imbalanced.
- We implemented a decision tree model that achieved an accuracy of 82% on the testing set.
- We performed correlation analysis between all columns and especially between the heart disease factors to find correlations between them and focus our study part.

- The correlation chart shows that the target variable “HeartDisease” has a moderate positive correlation with “Age” (0.23), “BMI” (0.04), “Diabetes” (0.09), “Smoking” (0.1), DiffWalking” (0.15),PhysicalActivity” (0.12), and “GeneralHealth” (0.2).
- It also shows that the target variable “HeartDisease” has a weak negative correlation with “Race” (-0.5) and a very weak negative correlation with “Gender” (-0.08).

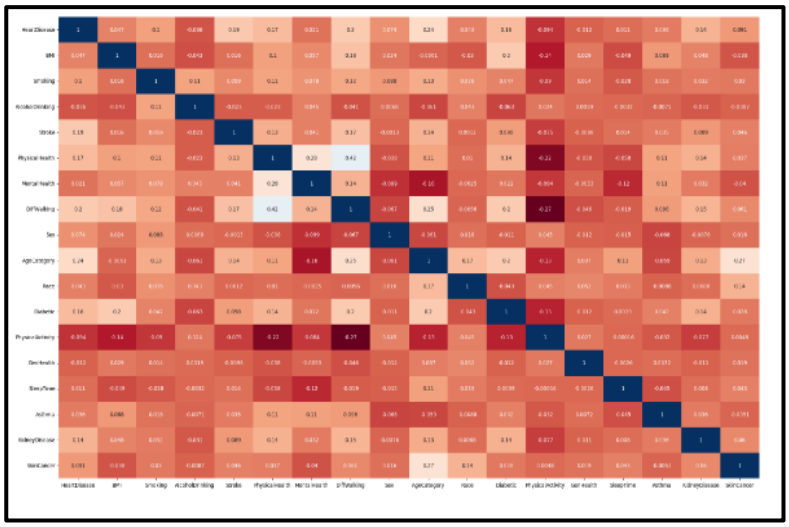

Figure 6: correlation Matrix

Predictive Modeling (Decision Tree):

- Our dataset is imbalanced, meaning that there are more respondents who do not have heart disease than those who do have heart disease. This can affect the performance and accuracy of the classifier models.
- To address the imbalance problem, the dataset is **split into training and testing sets** using **stratified sampling,** which ensures that the proportion of respondents with and without heart disease is similar in both sets. Then**, SMOTE (Synthetic Minority Oversampling Technique)** is applied to the training set to create synthetic samples of respondents with heart disease, so that the number of respondents with and without heart disease is equal in the training set.
- **A decision tree is** trained on the balanced training set using the scikit-learn library in Python. The decision tree uses the Gini impurity criterion to measure the quality of each split and the best split strategy to choose the best feature to split on at each node.

- The decision tree is evaluated on the testing set using various metrics, such as accuracy and precision. The decision tree achieves an **accuracy of 0.82,** a precision of 0.87, a recall of 0.75, and an F1-score of 0.30.
- The chart shows the feature importance of different health-related variables. BMI has the highest relative importance of around 0.35 while kidney disease has the lowest relative importance of around 0.05. Other variables include age category, sleep time, physical health, mental health, sex, smoking, race, physical activity, diabetes, difficulty walking, alcohol drinking, asthma, skin cancer and stroke.

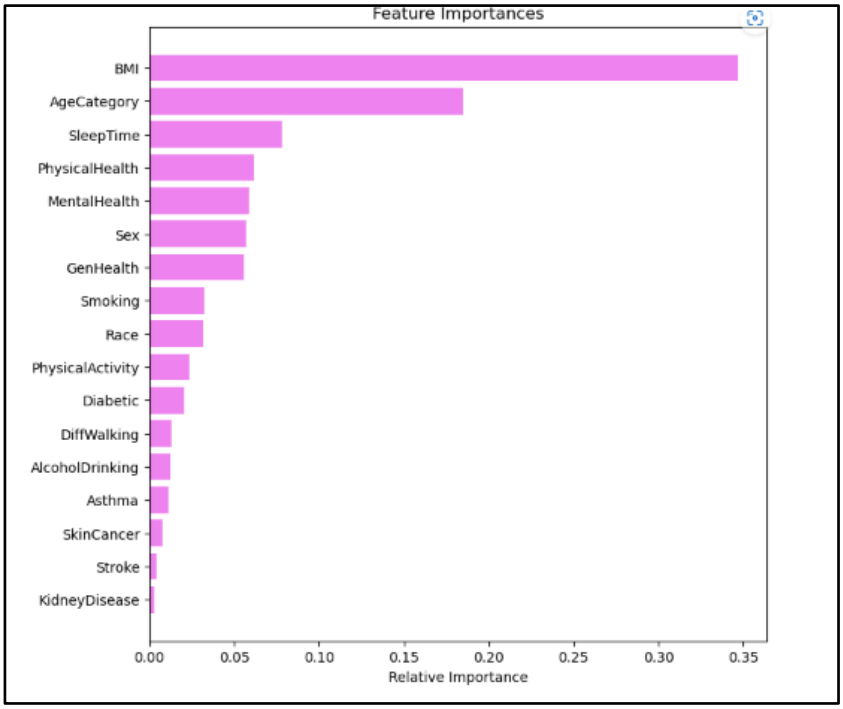

Figure 7: Decision tree result

Dashboard

Use Case:

The dashboard shows the trends and patterns of heart disease and factors affecting it in the USA. The primary use cases are to compare the heart disease risk across different races, ages, and sexes, to explore the impact of lifestyle factors such as walking and smoking on heart health, and to identify the areas that need more prevention and intervention. A user can interact with the dashboard by selecting a race (American, Asian, Black, Hispanic, or White), a sex (Male or Female), and an age category (18-24, 25-29, 30-34,… , 80+) from the filters, and hovering over the pie charts and bar charts to see more details.

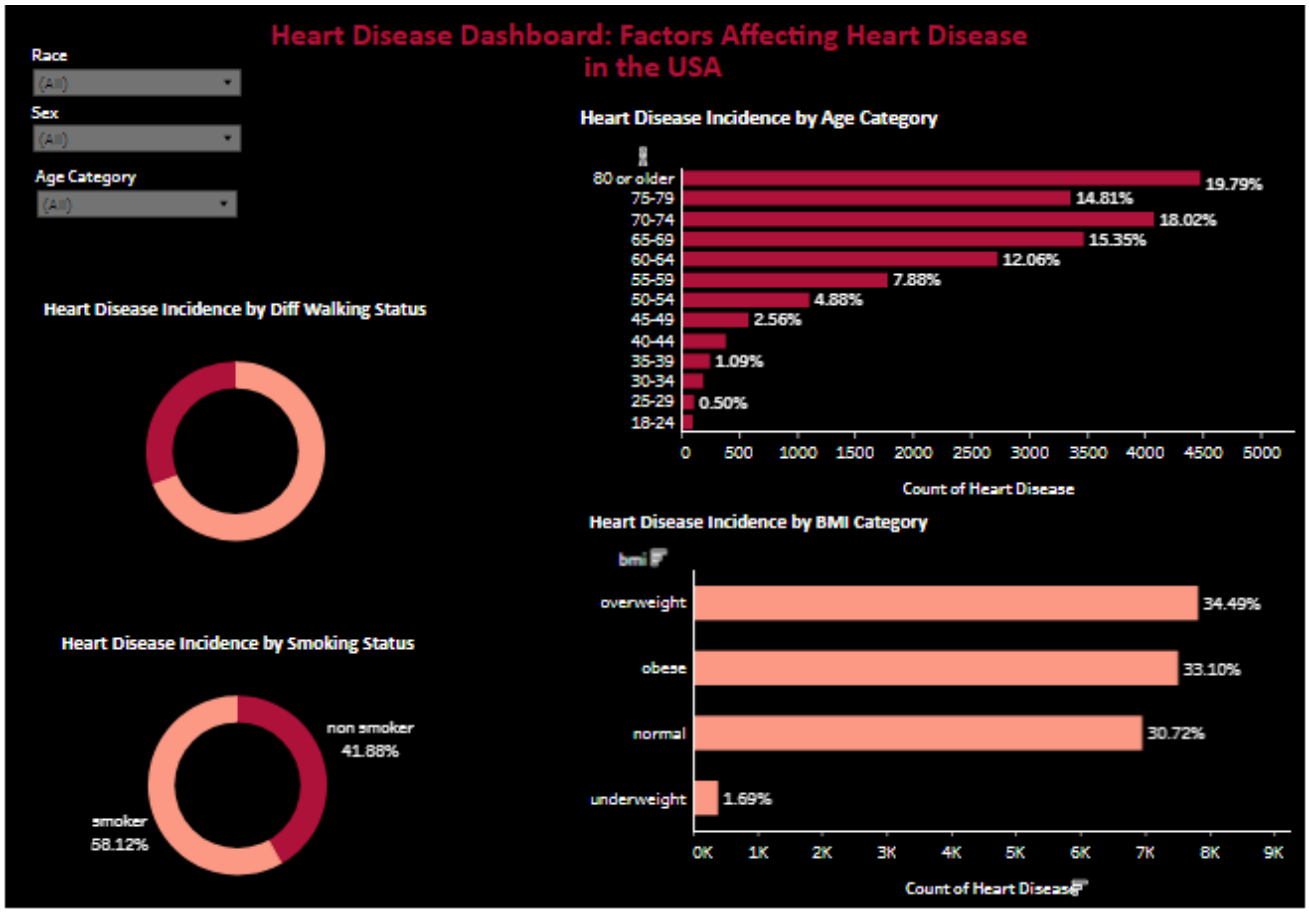

Figure 8 : Tableau Dashboard

Data Engineering:

- The data for our dashboard entitled “Key Indicators of Heart Disease” was obtained from Kaggle.
- The dataset contains data from the Behavioral Risk Factor Surveillance System (BRFSS), which is an annual telephone survey that collects data on the health status of U.S. residents.
- The raw data was processed and cleaned using **Python scripts.**
- The data was then loaded into **Tableau Desktop** where we created various charts, maps, and tables to visualize the heart disease trends and patterns.
- We chose to include pie charts to show the distribution of walking and smoking status among different groups, and bar charts to show the relationship between age category, BMI, and heart disease risk.
- We also added three **filters** to allow users to select a race, sex, and age category of interest.
- We organized our data into separate sheets for each chart and filter, and used calculated fields and parameters to create dynamic visualizations.
- We also considered the results of a **decision tree analysis that** showed that **age category and BMI** have **the most effect** on heart disease risk among other factors such as alcohol drinking, kidney disease, asthma, and skin cancer.

Conclusion & Future Work

Highlights

- The target variable “HeartDisease” is influenced by several factors, both demographic and health related.
- We found that gender, age, race, obesity, diabetes, physical activity, general health, mental health, alcohol consumption, smoking, stroke, Diffwalking, asthma, kidney disease, and skin cancer are all significant predictors of heart disease.
- We developed a decision model that achieved an accuracy of 82% on the testing set. Found that Age and BMI are the most influential indicators affecting heart disease risk, followed by smoking.
- Unhealthy habits like smoking and excessive alcohol consumption increase the risk by 2 to 3 times.
- pre-existing conditions like diabetes and kidney disease have positive correlations with heart disease risk. This means that they heighten the likelihood of heart disease and emphasize the need for caution and lifestyle changes. Healthcare providers and policymakers should focus on preventing and managing these conditions to reduce the incidence of heart disease.

Conclusion:

From the data analysis and prediction modeling on the dataset from Kaggle, we concluded that there are several factors that affect heart disease risk and that these factors interact with each other in complex ways. We developed a decision tree model that can accurately predict heart disease based on these factors and provide insights into the most important ones. We found that age and BMI are the most influential indicators affecting heart disease risk, followed by diabetes, smoking, and general health. We also found that gender, race, physical activity, mental health, alcohol consumption, stroke, walking difficulty, asthma, kidney disease, and skin cancer have significant effects on heart disease risk. We provided insights and recommendations for healthcare providers and policymakers to improve the prevention and management of heart disease based on our findings. We suggested that screening for heart disease should consider all relevant risk factors and not only the major ones. We also suggested that interventions should target the modifiable risk factors such as diabetes, smoking,

and alcohol consumption. We also suggested that public health policies should promote healthy lifestyles and behaviors such as physical activity, mental health, and regular check-ups.

Future work

- Find a dataset that contains more recent data on heart disease and its risk factors and update our prediction model accordingly.
- Find a dataset that covers other countries or regions besides the U.S. and compare the differences and similarities in heart disease risk and prevalence.
- Evaluate the impact of our prediction model and recommendations on the actual prevention and management of heart disease in real-world settings.